Frontiers in ***Cognitive Science***

8January 2014

Self-Organization of Early Vocal Development in Infants and Machines: The Role of Intrinsic Motivation

Clément Moulin-Frier [1,*], **Sao Mai Nguyen** [1] **and Pierre-Yves Oudeyer** [1]

[1] *Flowers Team, Inria / ENSTA-Paristech, France*

Correspondence*:
Clément Moulin-Frier
INRIA Bordeaux Sud-Ouest, 200 avenue de la Vieille Tour, 33 405 Talence Cedex, France, clement.moulinfrier@gmail.com

Research Topic

ABSTRACT

We bridge the gap between two issues in infant development: vocal development and intrinsic motivation. We propose and experimentally test the hypothesis that general mechanisms of intrinsically motivated spontaneous exploration, also called curiosity-driven learning, can self-organize developmental stages during early vocal learning. We introduce a computational model of intrinsically motivated vocal exploration, which allows the learner to autonomously structure its own vocal experiments, and thus its own learning schedule, through a drive to maximize competence progress. This model relies on a physical model of the vocal tract, the auditory system and the agent's motor control as well as vocalizations of social peers. We present computational experiments that show how such a mechanism can explain the adaptive transition from vocal self-exploration with little influence from the speech environment, to a later stage where vocal exploration becomes influenced by vocalizations of peers. Within the initial self-exploration phase, we show that a sequence of vocal production stages self-organizes, and shares properties with data from infant developmental psychology: the vocal learner first discovers how to control phonation, then focuses on vocal variations of unarticulated sounds, and finally automatically discovers and focuses on babbling with articulated proto-syllables. As the vocal learner becomes more proficient at producing complex sounds, imitating vocalizations of peers starts to provide high learning progress explaining an automatic shift from self-exploration to vocal imitation.

Keywords: Vocal Development, Intrinsic Motivation, Curiosity-driven learning, Imitation, Self-Organization, Interactive Learning, Goal Babbling.

1 INTRODUCTION

1.1 VOCAL DEVELOPMENT AND INTRINSIC MOTIVATION

Early on, babies seem to explore vocalizations as if it was a game in itself, as reported by Oller [1] who cites two studies from the 19th century:

> "[At] three months were heard, for the first time, the loud and high crowing sounds, uttered by the child sponteaneously, [...] the child seemed to take pleasure in making sounds." [2]
> "[He] first made the sound *mm* spontaneously by blowing noisily with closed lips. This amused [him] and was a discovery for [him]."[1] [3]

Such play with his vocal tract, where the baby discovers the sounds he can make, echoes other forms of body play, such as exploration of arm movements or how he can touch, grasp, mouth or throw objects. The concept of *intrinsic motivation* has been proposed in psychology to account for such spontaneous exploration [4, 5, 6, 7, 8]:

> "Intrinsic motivation is defined as the doing of an activity for its inherent satisfaction rather than for some separable consequence. When intrinsically motivated, a person is moved to act for the fun or challenge entailed rather than because of external products, pressures or reward." [6]

Intrinsic motivation refers to a mechanism pushing individuals to select and engage in activities for their own sake because they are inherently interesting (in opposition to *extrinsic motivation*, which refers to doing something because it leads to a separable outcome). A key idea of recent approaches to intrinsic motivation is that *learning progress* in sensorimotor activities can generate intrinsic rewards in and for itself, and drive such spontaneous exploration [8]. Learning progress refers to the infant's improvement of his predictions or control over activity they practice, which can also be described as reduction of uncertainty [9].

[1] We have changed the gender of the subject to a male in this quotation, in order to follow the convention of the present article. Throughout this paper, we will use "he" for an infant, "she" for a caregiver (e.g. the mother) and "it" for a learning agent (the model).

Although spontaneous vocal exploration is an identified phenomenon, occurring in the early stages of infant development, the specific mechanisms of such exploration and the role of intrinsic motivation for the *structuration* of early vocal development has not received much attention so far to our knowledge. We propose that mechanisms of intrinsically motivated spontaneous exploration, which we also refer to as curiosity-driven learning, play an important role in speech acquisition, by driving the infant to follow a self-organized developmental sequence which will allow him to progressively learn to control his vocal tract. This is to our knowledge a largely unexplored hypothesis. The goal of this article is to formalize in detail this hypothesis and study general properties of such mechanisms in computer experiments.

Several computational models of speech development, where speech acquisition is organized along a developmental pathway, have been elaborated so far. They have shown how such stage-like organization can ease the acquisition of complex realistic speech skills.

The DIVA model [10, 11], as well as Kröger's model [12], propose architectures partly inspired by neurolinguistics. They involve two learning phases. The first one is analogous to infant babbling and corresponds to semi-random articulator movements producing auditory and somatosensory feedbacks. This is used to tune the correspondences between representation maps within a neural network . In the second phase, the vocal learner is presented with external speech sounds analogous to an ambient language and learns how to produce them adequately. The Elija model [13] also distinguishes several learning phases. In the first phase of exploration, the agent is driven by a reward function, including intrinsic rewards such as sound salience and diversity, as well as articulatory effort. Various parameterizations of this reward function allows the model to produce vocalizations in line with Oller's vocal developmental stages of infants. In a subsequent phase, the sounds produced by the model attract the attention of a caregiver, providing an external reinforcement signal. Other models also use a reinforcement signal, either from human listeners (social reinforcement [14, 15]) or based on sound saliency (intrinsic reinforcement [16]), and show how this can influence a spiking neural network to produce canonical syllables. Such computational models of speech acquisition pre-determine the global ordering and timing of learning experiences, which amounts to preprograming the developmental sequence. Understanding how a vocal developmental sequence can be formed is still a major mystery to solve, and this article attempts a first step in this direction.

We build on recent models of skill learning in other modalities (e.g. locomotion or object manipulation), where it was shown that mechanisms of intrinsically motivated learning can self-organize developmental pathways, adaptively guiding exploration and learning in high-dimensional sensorimotor spaces, involving highly redundant and non-linear mappings [17, 18, 19, 8]. Such models concretely formalize concepts of intrinsic motivation described in the psychology literature into algorithmic architectures that can be experimented in computers and robots [20, 21, 22, 23]. Detailed discussions of the engineering aspects of such intrinsic motivation mechanisms, casted in the statistical framework of active learning, have been recently published and showed their algorithmic efficiency to learn sensorimotor coordination skills in redundant non-linear high-dimensional mappings [24, 18, 25].

Indeed, transposed in curiosity-driven learning machines [20, 21, 26, 27, 28, 29, 30] and robots [17, 18], these developmental mechanisms have been shown to yield highly efficient learning of inverse models in high-dimensional redundant sensorimotor spaces [18, 31]. These spaces share many mathematical properties with vocal spaces. Efficient versions of such mechanisms are based on the active choice of learning experiments that maximize learning *progress*, e.g. improvement of predictions or of competences to reach goals [20, 17, 22, 18, 25]. Such learning experiments are called "progress niches" [17].

Yet, beyond pure considerations of learning efficiency, exploration driven by intrinsic rewards measuring learning progress was also shown to self-organize structured developmental pathways, both behaviorally and cognitively. Indeed, such mechanisms automatically drive the system to explore and learn first easy skills, and then progressively explore skills of increasing complexity [17]. They have been shown to generate automatically behavioural and cognitive developmental structures and have been analyzed in relation to their similarities with infant development [17, 32, 33, 34]. For example, in the Playground Experiment, a curiosity-driven learning robot was shown to self-organize its own learning experiences into a sequence of behavioural and cognitive stages where it spontaneously acquired various affordances and skills of increasing complexity [17]. It was also shown how it could discover and focus on elementary vocal interaction with a peer as a spontaneous consequence of its general drive to explore situations where it can improve its predictions [33]. Focusing on vocal interactions was thus explained as a special case of focusing on an activity that provides learning progress (i.e. a particular progress niche). This therefore allowed to generate some novel hypotheses to explain infant development, from the behavioural [33], cognitive [32] or brain circuitry [35] perspectives (see [8] for a review on these novel perspectives). Intrinsically motivated spontaneous learning has also been combined with mechanisms of imitation learning within the SGIM-ACTS architecture, as detailed in [36]. In this model, formulated within the framework of strategic learning [37], a hierarchical active learning architecture allows an interactive learning agent to choose by itself when to explore autonomously, and when, what and who to imitate, based on measures of competence progress.

Although intrinsic motivation and socially guided learning have already been considered in computational models specifically studying speech acquisition, to our knowledge, they have so far been considered as two distinct learning phases with a hard-coded switch between them (e.g. [10, 11, 12, 13]). In other words, the existence of distinct developmental stages was presupposed in these models. In contrast, these distinct learning phases emerge from the Playground Experiment, even though only a simplistic vocal system was considered (only pitch and duration were controlled, and no physical model of the vocal tract was used; modeling of speech acquisition per se was not the focus of this study).

Our main contribution in this paper is to show how mechanisms of intrinsically motivated exploration applied on a realistic articulatory-auditory system self-organizes autonomously into coherent *vocal* developmental sequences. This follows the approach of our previous works [34, 38, 39], which were preliminary studies limited to vowel production and focusing only on autonomous learning, i.e. without considering a surrounding ambient language.

In such a conceptual framework, developmental structures are neither learnt from "tabula rasa" nor a pre-determined result of an innate "program": they self-organize out of the dynamic interaction between constrained cognitive mechanisms (including curiosity, learning, and

abstraction), the morphological properties of the body, and the physical and social environment which itself is constrained and ordered by the developmental level of the organism [17, 40]. Thus, the approach we take can be viewed as an instantiation of the concept of epigenesis, in the sense proposed by [41].

The study of such a dynamical systems approach, where curiosity-driven learning is an important force, can take ample advantage of computer modeling as a research tool. Here in particular, it can help to understand better the dynamics underlying early vocal development, and in particular understand what are the mechanisms which generate the developmental sequence(s) in vocal productions and capabilities observed in infants. In particular, it can help to understand what is the precise role of intrinsic motivation.

In the next sections of this introduction, we summarize properties of vocal development during the first year and describe the general principles of the computational model we study in this article.

1.2 DEVELOPMENT OF VOCALIZATIONS

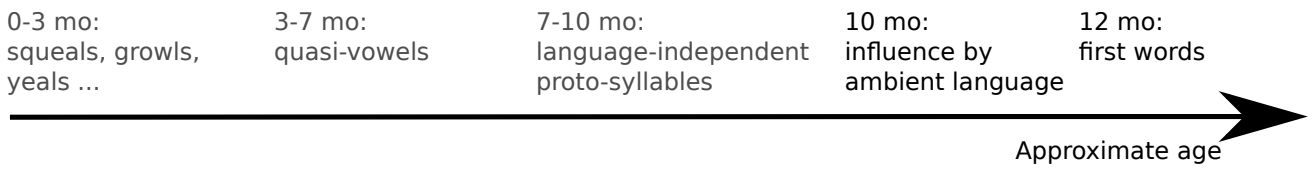

Figure 1. The first year of infant vocal development.

Despite inter-individual variations in infant vocal development (e.g. [42]), strong regularities in the global structuration of vocal development are identified [1, 43]. In this article, we adopt the view from Oller [1] as well as Kuhl [43]. **Figure 1** schematizes this vocal development during the first year of infant. It can be summarized as follows. First, until the age of approximately 3 months, an infant produces non-speech sounds like squeals, growls and yeals. During this period, he seems to learn to control infrastructural speech properties, e.g. phonation and primitive articulation [1]. Then, from 3 to 7 months, he begins to produce vowel-like sounds (or quasi-vowels) while he probably learns to control his vocal tract resonances. At 7 months, canonical babbling emerges where well-timed sequences of proto-syllables are mastered. But it is only around the age of 10 months that infant vocal productions become more influenced by the ambient language, leading to first word productions around 1 year of age.

Two features of this developmental sketch are particularly salient.

- Infants seem to first play with their vocal tracts in a relatively language-independent way, and then are progressively influenced by the ambient speech sounds.
- In the initial phase, when sounds produced by their peers influence little their vocalizations, infants seem to learn skills of increasing complexity: normal phonation, then quasi-vowels and finally proto-syllables. According to Oller [1], such a sequence displays a so-called natural, or logical hierarchy. For example, it is impossible to master quasi-vowel production without previously mastering normal phonation.

1.3 A COMPUTATIONAL MODEL OF VOCAL DEVELOPMENT

To articulate hypotheses about the possible roles of intrinsic motivation in the first year of vocal development, we build here a computational model of an intrinsically motivated vocalizing agent, in contact with vocalizations of peers. In the model, an individual speech learner has the following characteristics, described in detail in next sections:

- It embeds a realistic model of a human vocal tract: the articulatory synthesizer used in the DIVA model [44]. This model provides the way to produce sequences of vocal commands and to compute corresponding sequences of acoustic features, both in multi-dimensional continuous domains.
- It embeds a dynamical model for producing motions of the vocal tract, based on a an over-damped spring-mass model. This model describes dynamical aspects such as co-articulation in sequences of vocal targets.
- It is able to iteratively learn a probabilistic sensorimotor model of the articulatory-auditory relationships according to its own experience with the vocal tract model. Because the sensorimotor learning is iterative during the life time of the agent, it will first be inefficient at using this model for control, and then progresses by learning from its own experience.
- It is equipped with an intrinsically motivated exploration mechanism, which allows it to generate and select its own auditory goal sequences. Such mechanism includes a capability to empirically measure its own competence progress to reach sequences of goals. Then, an action selection system stochastically self-selects target goals that maximize competence progress.

- It is able to hear sounds of a simulated ambient language, and its intrinsic motivation system is also used to decide whether to self-explore self-generated auditory goals, or to try to emulate adult sounds. This choice is also based on a measure of competence progress for each strategy.

Then, we present experiments allowing us to study how the developmental structuration of early vocal exploration could be self-organized in an intrinsically motivated speech learner, under the influence of sounds in the environment and constrained by the physical properties of the sensorimotor system.

In a first series of experiments, we consider a speech learner who is not exposed to external speech sounds. This allows the study of the role of intrinsic motivation independently of any social influence. We show how a cognitive architecture for intrinsically motivated autonomous exploration (SAGG-RIAC, [18, 39]), applied to learning to control an articulatory synthesizer (i.e. a vocal tract model able to produce speech sounds from articulatory configurations), can self-organize coherent vocal developmental sequences. This work extends preliminary studies [34, 39, 38] through the use of a different vocal tract model and a more complex model of motion control dynamics with an overdamped spring-mass dynamical system, providing the agent with a more realistic and powerful mechanism to produce (un)articulated sounds.

In a second series of experiments, the speech learner is exposed to speech sounds from its environment. The cognitive architecture is extended to strategic interactive intrinsically motivated learning (SGIM-ACTS, [36]), where intrinsic motivation is also used by the learner to decide when to self-explore and when to try to imitate sounds in the environment. In the present study, we suppose that the sounds of the adult are directly imitable (we do not account for the pitch and formant differences between infants and adults for instance). We show that the system first focuses on self-exploration of vocalization. It later on shifts to vocal imitation, which then influences its vocal learning in ways that are specific to the speech environment. Yet, in this paper, we do not study the social interaction aspect of the teacher and, in particular, we do not model the behavior of the adult in response to the learner behavior.

Our aim is to study how important aspects of infant vocal development in the first year of life, described in the previous section, could be explained by the interaction between these building blocks: an intrinsic motivation system, a dynamic motor system associated to morphological and physiological constraints, an imitation system and a system for learning a sensorimotor model out of physical experiments. We will show that competence progress based autonomous exploration is able to provide a unified explanation for both the tendency to produce vocalizations of increasing complexity and the progressive influence of the ambient adult sounds. Imitating adult sounds becomes interesting for the speech learner only when basic speech production principles have been previously mastered. Contrarily to existing models of speech acquisition we described so far, our aim is not to reproduce infant vocalizations in a phonetically detailed manner, but rather to suggest an hypothesis about how a succession of distinct developmental stages can self-organize autonomously. Howard & Messum's model [13] for example, shows how distinct parameterizations of an intrinsic reward function can enable a vocal agent to discover several type of sounds coherent with observed developmental stages in infants. These parameterizations however, are hard-coded. In contrast, our model is not designed to reproduce precisely infant vocalizations within distinct vocalization stages, but rather to understand how the *transition* from one stage to another can be explained by a drive to maximize the competence progress to reach self-generated or ambient auditory goals. In consequence, the switch from self-generated auditory goals to the imitation of adult sounds is not hard-coded in our model, but emerges as a by-product of the drive to focus on progress niches.

2 MODEL

In this section, we describe the models that we use for the vocal tract and auditory signals. We describe the learning of the internal model of the sensorimotor mapping, and the intrinsic motivation mechanism which allows the learner to decide adaptively which vocalization to experiment at given moments during its development, and whether to do so through self-exploration or through imitation of external sounds.

2.1 SENSORIMOTOR SYSTEM

2.1.1 Vocal Tract and Auditory System Our computational model involves the articulatory synthesizer of the DIVA model described in [44][2] based on Maeda's model [45]. Without going into technical details, the model corresponds to a computational approximation of the general speech production principles illustrated in **Figure 2**. The model receives 13 articulatory parameters as input. The first 10 are from a principal component analysis (PCA) performed on sagittal contours of images of the vocal tract of a human speaker, allowing to reconstruct the sagittal contour of the vocal tract from a 10-dimensional vector. The effect of the 10 articulatory parameters from the PCA on the vocal tract shape is displayed **Figure 3**. In this study, we will only use the 7 first parameters (the effect of the others on the vocal tract shape is negligible), fixing the 3 last in the neutral position (value 0 in the software). Through an area function, associating sections of the vocal tract with their respective area, the model can compute the 3 first formants of the resulted signal if phonation occurs. Phonation is controlled through the 3 last parameters: glottal pressure controlling the intensity of the signal (from quiet to loud), voicing controlling the voice (from voiceless to voiced) and pitch controlling the tone (from low-pitched to high-pitched). It is then able to compute the formants of the signal (among other auditory and somato-sensory features) through the area function. In this study, we only use the glottal pressure and voicing parameters. In addition to the 7 articulatory parameters from the PCA, a vocal command is therefore defined by a 9-dimensional vector. From

[2] available online at `http://www.bu.edu/speechlab/software/diva-source-code`. DIVA is a complete neurocomputational model of speech acquisition, in which we only use the synthesizer computing the articulatory-to-auditory function.

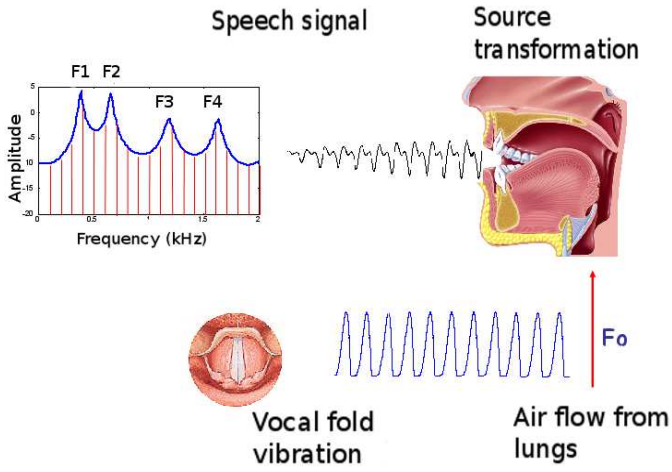

Figure 2. Speech production general principles. The vocal fold vibration by the lung air flow provides a source signal: a complex sound wave with fundamental frequency F_0. According to the vocal tract shape, acting as a resonator, the harmonics of the source fundamental frequency are selectively amplified or faded. The local maxima of the resulting spectrum are called the formants, ordered from the lower to the higher frequencies. They belong to the major features of speech perception.

the vocal command, the synthesizer computes the auditory and somatosensory consequences of the motor command, thus approximating the speech production principles of **Figure 2**.

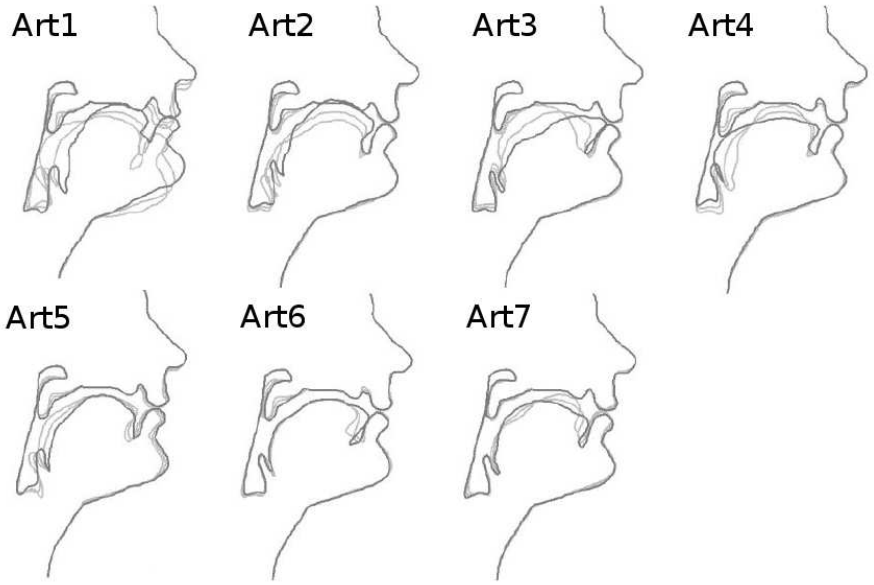

Figure 3. Articulatory dimensions controlling vocal tract shape (10 dimensions, from left to right and top to bottom), adapted from the documentation of the DIVA source code. Each subplot shows a sagittal contour of the vocal tract, where we can identify the nose and the lips on the right side. Bold contours correspond to a positive value of the articulatory parameter, the two thin contours are for a null (neutral position) and negative values. These dimensions globally correspond to the dimensions of movements of the human vocal tract articulators. For example, Art_1 mainly controls the jaw height, whereas Art_3 rather controls the tongue front-back position.

On the perception side of our model, we use the first two formants of the signal, $F1$ and $F2$, approximately scaled between -1 and 1. We also define a third parameter I which measures the intensity (or phonation level) of the auditory outcome. I is supposed to be 0 when the agent perceives no sound, and 1 when it perceives a sound. Technically, $I = 1$ if and only if two conditions are checked: (1) both pressure and

voicing parameters are above a fixed threshold (null value) and (2) the vocal tract is not closed (i.e. the area function is positive everywhere). In human speech indeed, the formants are not measurable when phonation is under a certain threshold. We model this by setting that when $I = 0$, the formants do not exist anymore and are set to 0. This drastic simplification is yet arguable in term of realism, but what we want to model here is the fact that no control of the formant values can be learnt when no phonation occurs.

2.1.2 Dynamical properties Speech production and perception are dynamical processes and the principles of **Figure 2** have to be extended with this respect. Humans control their vocal tract by variations in muscle activations during a vocalization, modulating the produced sound in a complex way. Closure or opening movements during a particular vocalization, coupled with variations in phonation level, are able to generate a wide variety of modulated sounds. We thus define a vocalization as a trajectory of the 9 motor parameters over time, lasting 800 milliseconds, from which the articulatory synthesizer is able to compute the corresponding trajectories in the auditory space (i.e. trajectories in the 3-dimensional space of $F1$, $F2$ and I). The agent is able to control this trajectory by setting 2 commands for each articulator: one from 0 to $250ms$, the other one from 250 to $800ms$. Then, the motor system is modeled as an overdamped spring-mass system driven by the following second-order dynamical equation:

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2(x - m) = 0, \tag{1}$$

where x is a motor parameter, and m is the command for that motor parameter. ζ is set to 1.01, ensuring that the system is overdamped (no oscillation), and ω_0 to $\frac{2\pi}{0.8}$ (0.8 being the duration of the vocalization in seconds). Thus, the agent's policy for a vocalization is defined by two vectors m_1 and m_2 (one for each command) of 9 real values each (one for each motor parameter). The policy space is 18-dimensional. The first command is applied for the beginning of the vocalization to $250ms$, the second one from $250ms$ to $800ms$.

Figure 4A illustrates the process by showing a typical syllabic vocalization. In this illustrative example, the controlled articulators are the first and third articulators of **Figure 3** (roughly controlling the jaw height and the tongue front/back dimensions), as well as pressure and voicing. The two last ones are set to 0.5 and 0.7 respectively, for both commands, to allow phonation to occur. The "jaw parameter" ($art1$ on the figure) is set to 2.0 (jaw closed) for the first command and to -3.0 for the second one (jaw open). We observe that these commands, quite far from the neutral position, are not completely reached by the motor system. This is due to the particular dynamics of the system, defined with ζ and ω_0 in the dynamical system. For the third articulator ($art3$), the commands are both at 2.0. We observe that, whereas the value 2.0 cannot be achieved completely at $250ms$, it can however be reached before the end of the vocalization.

This motor system implies interaction between the two commands, i.e. a form of co-articulation. Indeed, a given motor configuration may sometimes be harder to reach if it is set as the first command, because time allocated to reach the first command is less than for the second command. Reversely, some movements may be harder to control in the second command because the final articulator positions will depend both on the first and the second commands (e.g., it is harder to reach the value -3.0 for the second command if the first command is set to 2.0, than if the first command is set to -3.0, as seen in the example of **Figure 4**).

These characteristics are the results of modeling speech production as a damped spring-mass system (Eq. 1), which is a common practice in the literature [46, 47, 13].

Figure 4B shows the resulting vocal tract shape at the end of the 2 commands (i.e. at $250ms$ and at $800ms$). We observe that the vocal tract is closed at the end of the first command, open at the end of the second one.

Figure 4C shows the resulting sound. We observe that there is no sound during vocal tract closure.

Figure 4D shows the resulting trajectories of auditory parameters. In our experiments, we model the auditory perception of the agent of its own vocalization as the mean value of each parameter I, F1 and F2 in two different time windows lasting $150ms$: the first one from 250 to $400ms$, the second one from 650 to $800ms$. The auditory representation of a vocalization is therefore a 6-dimensional vector $(I(1), I(2), F1(1), F1(2), F2(1), F2(2))$. Perceived auditory values are represented by circles on **Figure 4D**. Note that the agent does not have any perception of what happens before $250ms$, and that $I(1)$ and $I(2)$ can take continuous values in $[0, 1]$ due to the averaging in a given perception time window. We will refer to the perceived "phone" of a given command for the perception occurring around the end of that command, although such an association will not be assumed in the internal sensorimotor model of the agent. Indeed, this sensorimotor system has the interesting property that the perceptions in both time windows depend on both motor commands. In the example of **Figure 4**, the perception for the first command, i.e. the mean auditory values between 250 and $400ms$, would not be the same if the second motor command did not cause the vocal tract opening.

2.1.3 Vocalization classification We define three types of phones, according to the value of I for a given command. In this description, we use common concepts like vowels or consonants to make an analogy with the human types of phones, although this analogy is limited.

- Those where $I > 0.9$: , i.e. phonation occurs during almost all the $150ms$ of perception around the end of the command. We call them *Vowels* (V).
- Those where $I < 0.1$, i.e. there is almost no phonation during the $150ms$ of perception around the end of the command. We call them *None* (N).
- Those where $0.1 < I < 0.9$, i.e. phonation occurs partially during the 0.15s of perception around the end of the command. This means that the phonation level I has switched during that period. This can be due either to a closure or opening of the vocal tract, or to variations in the pressure and voicing parameters. We call them *Consonants* (C), although they are sometimes more comparable to a sort of prosody (when due to a variation in the phonation level).

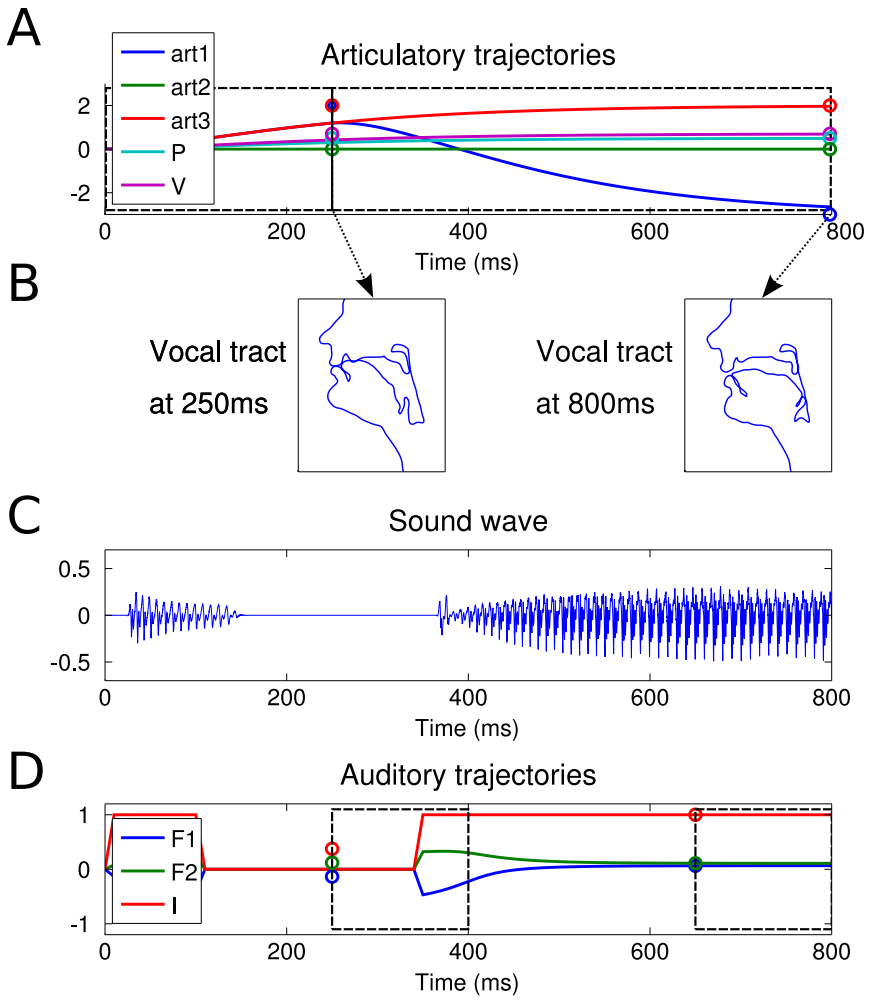

Figure 4. An illustrative vocalization example. A) Articularory trajectories of 5 articulators during the 800ms of the vocalization (4 articulators, from $art4$ to $art7$ are not plotted for the sake of readability but display the same trajectory as $art2$). Circles at 250 and 800ms represents the values of the first and second commands, respectively, for each trajectory. The first commands are active from 0 to 250ms and second ones from 250 to 800ms, as represented by dotted black boxes. The trajectories are computed by the second order dynamical equation (1), starting in a neutral position (all articulators set to 0). B) Resulting vocal tract shapes at the end of each command, i.e. at 250 and 800ms. Each subplot displays a sagittal view with the nose and the lips on the left side. The tongue is therefore to the right of the lower lip. C) Sound wave resulting from the vocalization. D) Trajectories of the 3 auditory parameters, the intensity I and the two first formants $F1$ and $F2$. Dotted black boxes represent the two perception time windows. The agent perceives the mean value of the auditory parameters in each time window, represented by the circles at 250 and 650ms.

This classification will be used as a tool for the analysis of the results in section 3, but is never known by the agent (which only has access to the values of I, $F1$ and $F2$).

Thus, each vocalization produced by the agent, belongs to the combination of 2 of these 3 types (because a vocalization corresponds to 2 commands), i.e. there are $3^2 = 9$ types of vocalizations: VV, VN, VC, NV, NN, NC, CV, CN, CC. An example of each type is given in the supplementary data, section 4.1.

Then, we suggest to group these 9 types into 3 classes.

- The class *No Phonation* contains only NN: the agent has not produced an audible sound. This is due either to the fact the pressure and voicing motor variables have never been sufficiently high (not both positive, as explained in the description of the motor system) during the two $150ms$ perception periods, or that the vocal tract was totally closed.

- The class *Unarticulated* contains VN, NV, CN, NC: the vocalization is not well-formed. Either the first or the second command produces a phone of type *None* ($I < 0.1$, see above).
- The class *Articulated* contains CV, VC, VV and CC: the vocalization is well-formed, in the sense that there is no *None* phone. Phonation is modulated in most cases (i.e. except in the rare case where the two commands of a VV are very similar). Note that according to the definition of *consonants*, phonation necessarily occurs in both the perception time windows (see **Figure 14** in the supplementary data).

It is important to note that the auditory values of these vocalization classes span subspaces of increasing complexity. Indeed, whereas various articulatory configurations belong to the *No Phonation* class, their associated auditory values are always null, inducing a 0-dimensional auditory subspace (i.e. a point). Regarding the *Unarticulated* class, the associated auditory values span a 3-dimensional subspace because at least one command produces a phone of type *None* (i.e. the corresponding auditory values are null). Finally, in the *Articulated* classes, the auditory values span the entire 6-dimensional auditory space. These properties will have important consequences for the learning of a sensorimotor model by the agent, as we will see.

2.2 INTERNAL SENSORIMOTOR MODEL

The sensorimotor internal model and the intrinsic motivation system which follow were firstly described in conference papers [39, 38] in a more general context where the goal was to compare various exploration strategies. In this paper, we use the active goal exploration strategy – analog to the SAGG-RIAC algorithm in [18, 31].

During its life time, the agent iteratively updates an internal sensorimotor model by observing the auditory results of its vocal experiments. We denote motor commands M and sensory perceptions S. We call $f : M \rightarrow S$ the unknown function defining the physical properties of the environment (including the agent's body). When the agent produces a motor command $m \in M$, it then perceives $s = f(m) \in S$, modulo an environmental noise and sensorimotor constraints. In the sensorimotor system defined in the previous section, M is 18-dimensional and S is 6-dimensional. f corresponds to the transformation defined section 2.1 and illustrated **Figure 4**, and has a Gaussian noise with a standard deviation of 0.01. By collecting (m, s) pairs through vocal experiments, the agent learns the joint probability distribution defined over the entire sensorimotor space SM (therefore 24-dimensional). This distribution is encoded in a Gaussian Mixture Model (GMM) of 28 components, i.e. a weighted sum of 28 multivariate normal distributions[3]. Let us note G_{SM} this GMM. It is learnt using an online version of the Expectation-Maximization (EM) algorithm [48] proposed by [49] where incoming data are considered incrementally. Each update is executed once each sm_step (=400) vocalizations are collected. G_{SM} is thus refined incrementally during the agent life, updating each time a number sm_step of new (m, s) pairs are collected. Moreover, we adapted this online version of EM to introduce a *learning rate* parameter α which decreases logarithmically from 0.1 to 0.01 over time. α allows to set the relative weight of the new learning data with respect to the old ones.

This GMM internal model is used to solve the inverse problem of inferring motor commands $m \in M$ that allow the learner to reach a given auditory goal $s_g \in S$. From this sensorimotor model G_{SM}, the agent can compute the distribution of the motor variables knowing a given auditory goal to reach s_g, noted $G_{SM}(M \mid s_g)$. This is done by Bayesian inference on the joint distribution, and results in a new GMM over the motor variables M (see e.g. [49]), from which the agent can sample configurations in M.

The whole process is illustrated **Figure 5**, on a toy example with mono-dimensional M and S. Given the current state of the sensorimotor model, the agent tries to achieve three goals, $s_1 = -9$, $s_2 = 0$, and $s_3 = 8$, i.e. three points in S (how the agent is going to self-generate such goals with intrinsic motivation will be explained below). At the beginning of the life time, the model is very poor at finding a good solution because the GMM is trained with only a few data, not necessarily concentrated in the regions useful to achieve the goals. For example, at $t = 500$, the agent is only able to correctly reach $s_2 = 0$ but is inefficient at reaching $s_1 = -9$ and $s_3 = 8$, as shown by the distributions over S in the top left corner (rotated 90 degrees anti-clockwise) . Then it becomes better and better while the agent produces new vocalizations, covering a larger part of the sensorimotor space: at $t = 1500$, the agent is able to reach the three goals.

The sensorimotor system we specified in the previous section, however, involves a 24-dimensional sensorimotor space (18 articularory dimensions and 6 auditory ones). Moreover, as we have already noted, the three vocalization classes we defined (*No Phonation, Unarticulated* and *Articulated*) span subspaces of the 6-dimensional auditory space with increasing dimensionality. Learning an inverse model using GMMs with a fixed number of Gaussians is harder, i.e. requires more sensorimotor experiments, as the spanned auditory subspace is of higher dimensionality. Although we do not provide mathematical arguments to this claim in this paper, it seems clear that learning an inverse model to produce *No Phonation* requires fewer learning data than learning an inverse model to produce various *Articulated* vocalizations, because the range of sensory effect is much larger in the second case.

2.3 INTRINSICALLY MOTIVATED ACTIVE EXPLORATION

In order to provide training data to the sensorimotor model we just described, the agent autonomously and adaptively decides which vocal experiments to make. The key idea is to self-generate and choose goals for which the learner predicts that experiments to reach these goals will lead to maximal competence progress.

The specific model we use in the first series of experiments (section 3.1) is a probabilistic version of the SAGG-RIAC architecture [18, 31]. This architecture was itself derived as a functional model [22, 8] of theories in psychology [4, 5, 6, 7] which describe spontaneous

[3] We empirically chose a number of components which is a suitable trade-off between learning capacity and computational complexity.

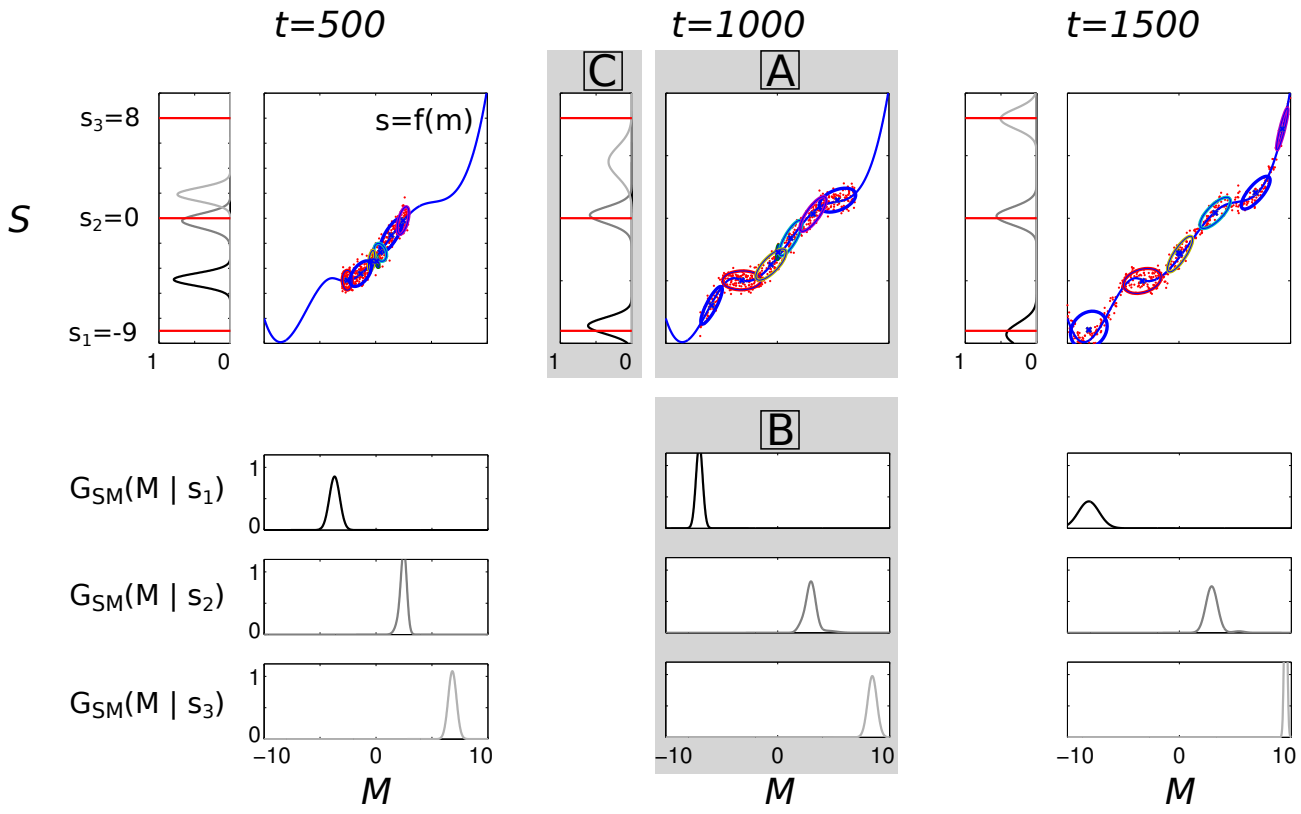

Figure 5. Illustration of incremental learning and inference in the sensorimotor model in a toy 2-dimensional sensorimotor space. The figure has three columns, corresponding to the state of a learning agent after 500, 1000 and 1500 sensorimotor experiments ($t = 500, 1000, 1500$). Each column is divided in three panels A, B and C, as indicated in the middle column (boxed letters in gray panels). X-axis (M space) and y-axis (S space) of A are shared by B and C, respectively. A) The unknown function $s = f(m)$ is represented by the blue curve. The red points are the sensorimotor experiments made at this stage (i.e. until the corresponding time index t): when m is produced, $s = f(m) + \epsilon$ is perceived, where ϵ is here a Gaussian noise with a standard deviation of 0.5. The ellipses represent the state of G_{SM} learned from the sensorimotor experiments, which is here a GMM with 6 components (each ellipse represents a 2D Gaussian). B) The three vertically-aligned plots show the motor distributions $G_{SM}(M \mid s_g)$ for 3 different goals, $s_1 = -9.0$ (top), $s_2 = 0.0$ (middle), and $s_3 = 8.0$ (bottom), in each of three columns (i.e. at the three time indexes). They are inferred from G_{SM} in A using Bayesian inference. C) The probability distributions on S (rotated 90 degrees anti-clockwise) resulting from sampling motor configurations according to $G_{SM}(M \mid s_g)$, to reach the three goals s_1, s_2, and s_3, the shade of grey of each one corresponding to that used in B: this means for example that, at a given time index t, producing motor commands according to the distribution $G_{SM}(M \mid s_3)$ (panel B, bottom) will result in sensory consequences following the darker distribution in panel C. The three considered goals s_1, s_2 and s_3 are represented by the three horizontal red lines, which are the same in the three columns. The distributions in C thus reflect how the learner is able to reach one of the three considered goals using the current state of its sensorimotor model: we observe that at $t = 500$, it can only reach $s_2 = 0$; at $t = 1000$, it can also reach $s_1 = -9$ and at $t = 1500$ it can reach those three goals.

exploration and curiosity in humans. It combines two principles: 1) goal babbling, also called goal exploration; 2) active learning driven by the maximization of empirically measured learning progress (which corresponds to the active goal strategy in [39, 38]). In practice, the learner self-generates its own auditory goals in the sensory space S. One goal is here a sequence of two auditory targets encoded in a 6-dimensional vector $s_g = (I(1), I(2), F1(1), F1(2), F2(1), F2(2))$ (see section 2.1). For each goal, it uses the current sensorimotor estimation to infer a motor program $m \in M$ in order to reach that goal. Through the sensorimotor system, this produces a vocalization and the agent perceives the auditory outcome $s \in S$, hence a new (m, s) training data. Goals are selected stochastically so as to maximize the expected competence progress (i.e. the learner is interested in goals where it predicts it can improve maximally its competence to reach them at a particular moment of its development). This allows the learner to avoid spending too much time on unreachable or trivial goals, and progressively explore self-generated goals/tasks of increasing complexity. As a consequence, the learner self-explores and learns only sub-parts of the sensorimotor space that are sufficient for reachable goals: this allows to leverage the redundancy of these spaces by building dense tubes of learning data only where it is necessary for control.

We define the competence c associated to a particular experiment (m, s) to reach the goal s_g as $c = comp(s_g, s) = e^{-\|s_g - s\|}$. This measure is in $[0, 1]$ and exponentially increases towards 1 when the Euclidean distance between the goal and the actual realization $s = f(m) + \epsilon$ tends to 0.

The measure of competence progress uses another GMM, G_{IM}, learnt using the classical version of EM on the recent goals and their associated competences. This GMM provides an interest distribution $G_{IM}(S)$ used to sample goals in the auditory space S maximizing the competence progress in the recent sensorimotor experiments of the agent. This was firstly formalized in [39, 38]. In this paper, we provide a graphical explanation of the process in **Figure 6**.

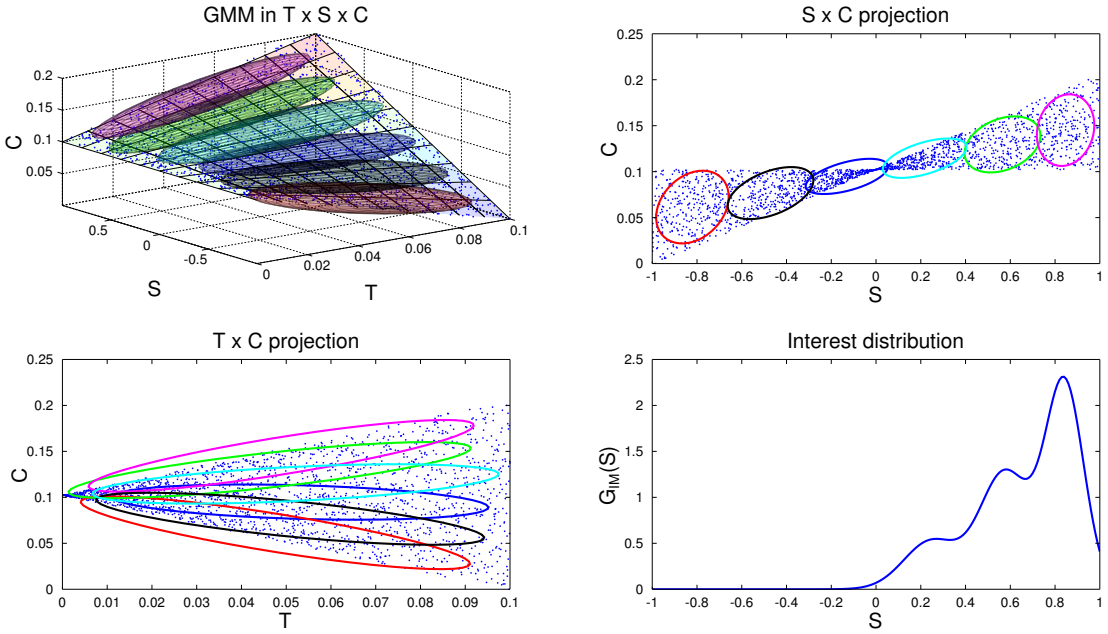

Figure 6. Illustration of interest distribution computation. Top-left: the recent history of competences of the agent, corresponding to blue points in the space $T \times S \times C$, where T is the space of recent time indexes (in $\mathbb{R}^+$), S the space of recently chosen goals s_g (mono-dimensional in this toy example) and C the space of recent competences of reaching those goals (in $\mathbb{R}^+$). For the sake of the illustration, the competence variations over time are here hand-defined (surf surface) and proportional to the values in S (increases for positive values, decreases for negative values). We train a GMM of 6 components, G_{IM}, to learn the joint distribution over $T \times S \times C$, represented by the six 3D ellipses. Projections of these ellipses are shown in 2D spaces $S \times C$ and $T \times C$ in the top-right and bottom-left plots. To reflect the competence progress in this dataset, we then bias the weight of each Gaussian to favor those which display a higher competence progress, that we measure as the covariance between time and competence for each Gaussian (in the example the purple ellipse shows the higher covariance in the bottom-left plot). We weight the Gaussians with a negative covariance between time T and competence C (blue, black and red ellipses) with a negligible factor, such that they do not contribute to the mixture. Using Bayesian inference in this biased GMM, we finally compute the distribution over the goal space S, $G_{IM}(S)$, thus favoring regions of S displaying the highest competence progress (bottom-right).

Following all the previous definitions, we now consider that the agent possesses the following abilities:

- Producing a complex vocalization, sequencing two motor commands interpolated in a dynamical system. It is encoded by a 18-dimensional motor configuration $m \in M$.
- Perceiving the 6-dimensional auditory consequence $s = f(m) + \epsilon \in S$, computed by an articularory synthesizer. f is unknown to the agent.
- Iteratively learning a sensorimotor model from lots of (m, s) pairs it collects by vocalizing through time. It is encoded in a GMM G_{SM} over the 24-dimensional sensorimotor space $M \times S$.
- Controling its vocal tract to achieve a particular goal s_g. This is done by computing $G_{SM}(M \mid s_g)$, the distribution over the motor space M knowing a goal to achieve s_g.
- Actively choosing goals to reach in the sensory space S by learning an interest model G_{IM} in the recent history of experiences. By sampling in the interest distribution $G_{IM}(S)$, the agent favors goals in regions of S which maximizes the competence progress.

This agent is thus able to act at two different levels. At a high level, it chooses auditory goals to reach according to its interest model G_{IM} maximizing the competence progress. At a lower level, it attempts to reach those goals using Bayesian inference over its sensorimotor model G_{SM}, and incrementally refines this latter with its new experiences. The combination of both levels results in a self-exploration algorithm (**Algorithm 1**).

Algorithm 1 Self-exploration with active goal babbling (stochastic SAGG-RIAC architecture).

1: initialise G_{SM} and G_{IM}
2: **while** true **do**
3: $\quad s_g \sim G_{IM}(S)$
4: $\quad m \sim G_{SM}(M \mid s_g)$
5: $\quad s = f(m) + \epsilon$
6: $\quad c = comp(s_g, s)$
7: $\quad update(G_{SM}, (m, s))$
8: $\quad update(G_{IM}, (s_g, c))$
9: **end while**

The agent starts in line 1 with no experience in vocalizing. Both GMMs have to be initialized in order to be used. To do this, the agent acquires a first set of (m, s) pairs, by sampling in M around the neutral values of the articulators (see **Figure 3**). Regarding the pressure and voicing motor parameters, we consider that the neutral value is at -0.25, which leads to *no phonation* (recall that both these parameters have to be positive for phonation to occur, section 2.1). This models the fact that the agent does not phonate in its neutral configuration, and has at least to raise the pressure and voicing parameters to be able do do it. The agent then executes this first set of motor configurations (mostly not phonatory), observes the sensory consequences, and initialises G_{SM} with the corresponding (m, s) pairs using incremental EM. G_{IM} is initialised by setting the interest distribution $G_{IM}(S)$ to the distributions of the sounds it just produced with this first set of experiences. Thus, at the first iteration of the algorithm, the agent tries to achieve auditory goals corresponding to the sounds it produced during the initialisation phase. Then, in the subsequent iterations, the interest distribution $G_{IM}(S)$ reflects the competence progress measure, and is computed as explained above.

Line 3, the agent thus selects stochastically $s_g \in S$ with high interest values. Then it uses $G_{SM}(M \mid s_g)$ to sample a vocalization $m \in M$ to reach s_g (line 4). The execution of m will actually produce an auditory outcome s (line 5), and a competence measure to reach the goal, $c = comp(s_g, s)$, is computed (line 6). This allows it to update the sensorimotor model G_{SM} with the new (m, s) pairs (line 7). Finally, it updates the interest model G_{IM} (line 8) with the competence c to reach s_g

Algorithm 1 will be run and the results analyzed in section 3.1.

2.4 SOCIAL (OR IMITATION) SYSTEM

In language acquisition and vocalization, the social environment plays naturally an important role. Thus we consider an active speech learner that not only can self-explore its sensorimotor space, but can also learn by imitation. In a second series of experiments (section 3.2), we extend the previous model by integrating the previous learning algorithm in the SGIM-ACTS architecture, which has been proposed in [36].

We consider here that the learning agent can use one of two learning strategies, which it chooses adaptively:

- explore autonomously with intrinsically motivated goal babbling, as described previously,
- or explore with imitation learning. We distinguish mimicry, in which the learner copies the policies of others without an appreciation of their purpose, from emulation, where the observer witnesses someone producing an outcome, but then employs its own policy repertoire to reproduce the outcome, as formalized in [50, 51, 52, 53]. As the learner a priori can not observe the vocal tract of the demonstrator, it can only emulate the demonstrator by trying to reproduce the auditory outcome observed, by using its own means, finding its own policy to reproduce the outcome. We consider that the demonstrator (the social peer) has a finite set of auditory outcomes, and every time the learner chooses to learn by social guidance, it chooses at random an auditory outcome among the set to emulate.

The learner can monitor the competence progress resulting from using each of the strategies. This measure is used to decide which strategy is the best progress niche at a given moment: a strategy is chosen with a probability directly depending on its associated expected competence progress. Thus, competence progress is used at two hierarchical levels of active learning, forming what is called strategic learning [37]: at the higher-level, it is used to decide when to explore autonomously, and when to imitate; at the lower-level, if self-exploration is selected, it is used to decide which goal to self-explore (as in the previous model). Since competence progress is a non-stationary measure and is continuously re-evaluated, the individual *learns* to choose both the strategy $str \in \{autonomous\_exploration, social\_guidance\}$ and the auditory goals $s_g \in S$ to target, by choosing which combination enables highest competence progress.

For the particular implementation of SGIM-ACTS of this paper, we use the same formalism and implementation as in **Algorithm 1** and consider that the strategy is another choice made by the agent. This leads to **Algorithm 2**, where the interest model G_{IM} now learns an interest distribution as in section 2.3. The difference is that the space of interest is now the union of the strategy space $\{autonomous\_exploration, social\_guidance\}$ and the auditory space S. We call $StrS$ this new space $StrS =$

$\{autonomous\_exploration, social\_guidance\} \times S$. Hence G_{IM} is a distribution over $StrS$ (**Algorithm 2**, line 3). If the self-exploration strategy is chosen ($str = autonomous\_exploration$), the agent acts as in Algorithm 1. If the social guidance strategy is chosen ($str = social\_guidance$, line 4), the learner then emulates an auditory demonstration $s_g \in S$ chosen randomly among the demonstration set of adult sounds (line 5), overwriting s_g of line 3. It then uses its sensorimotor model G_{SM} to choose a vocalization $m \in M$ to reach s_g, by drawing according to the distribution $G_{SM}(M \mid s_g)$ (line 7), as in the self-exploration strategy. The execution of m will produce an auditory outcome s (line 8), from which it updates its models G_{IM} and G_{SM} (lines 10 and 11).

Algorithm 2 Strategic active exploration (active goal babbling and imitation with stochastic SGIM-ACTS architecture).

1: Initialise G_{SM} and G_{IM}
2: **while** true **do**
3: $(str, s_g) \sim G_{IM}(StrS)$
4: **if** $(str = social\_guidance)$ **then**
5: $s_g \leftarrow$ random auditory demonstration from the ambient language
6: **end if**
7: $m \sim G_{SM}(M \mid s_g)$
8: $s = f(m) + \epsilon$
9: $c = comp(s_g, s)$
10: $update(G_{SM}, (m, s))$
11: $update(G_{IM}, (str, s_g, c))$
12: **end while**

Thus, this new exploration algorithm is augmented with yet another level of learning, allowing to choose between different exploration strategies. This strategy choice moreover uses the same mechanism as the choice of auditory goals, by means of the interest model G_{IM}.

Algorithm 2 will be run and the results analyzed in section 3.2.

3 RESULTS

The results of our experiments are presented in this section. We first run experiments where our agent learns in a pure self-exploration mode (**Algorithm 1**), without any social environment or sounds to imitate. In a second time, we introduce an auditory environment to study the influence of ambient language (**Algorithm 2**).

3.1 EMERGENCE OF DEVELOPMENTAL SEQUENCES IN AUTONOMOUS VOCAL EXPLORATION

We ran 9 independent simulations of **Algorithm 1** with the same parameters but different random seeds, of 240.000 vocalizations each[4]. Most of these 9 simulations display the formation of a developmental sequence, as we will see. Before describing the regularities and variations observed in this set of simulations, let us first analyse a particular one where the developmental sequence is clearly observable. **Figure 7** exhibits such a simulation. We observe three clear developmental stages, i.e three relatively homogeneous phases with rather sharp transitions. These stages are not pre-programmed, but emerge from the interaction of the vocal productions of the sensorimotor system, learning within the sensorimotor model, and the active choice of goals by intrinsically motivated active exploration. First (until $\simeq$ 30.000 vocalizations), the agent produces mainly motor commands which results in *no phonation* or in *unarticulated* vocalizations (in the sense of the classes defined section 2.1.3). Second (until $\simeq$ 150.000 vocalizations), phonation almost always occurs, but the vocalizations are mostly *unarticulated*. Third, it produces mainly *articulated* vocalizations.

The visualisation of the developmental sequence of the 9 independent simulations, provided **Figure 15** in the supplementary data, shows important interindividual variations whereas initial conditions are statistically similar due to initialisation in line 1 of **Algorithm 1**. These variations can be understood through the interaction of the sensorimotor system f, the internal sensorimotor model G_{SM} and the interest model G_{IM}, resulting in a complex dynamical system where observed developmental sequences are particular attractors (see e.g. [54, 55]). Moreover the sensorimotor and the interest models are probabilistic, thus inducing a non-negligible source of variability all along a particular simulation. Another factor is that using an online learning process on a GMM can result in a sort of forgetting, leading sometimes to the re-exploration of previously learnt parts of the sensorimotor space[5]. However, the sequence *No phonation* $\rightarrow$ *Unarticulated* $\rightarrow$ *Articulated* appears as a global tendency, as shown in **Table 1**. We observe that despite variations, most simulations begin with a mix of *no phonation* and *unarticulated* vocalizations, then mainly produce *unarticulated* vocalizations, and often end up with *articulated* vocalizations. An analogy can be made with human phonological systems, which are all different in the details but display strong statistical tendencies [56, 57, 58, 59].

[4] Each simulation involves several hours of computing on a desktop computer, due to the complexity of **Algorithm 1**, in particular in the Bayesian inference and update procedures.

[5] This is why we limited the simulations to 240.000 vocalizations each, in order to avoid this unwanted effect of forgetting. However, the fact that the system is able to adaptively re-explore sensorimotor regions that have been forgotten is an interesting feature of curiosity-driven learning.

This suggests that the agent explores its sensorimotor space by producing vocalizations of increasing complexity. The class *no phonation* is indeed the easiest to learn to produce for two reasons: the rest positions of the pressure and voicing motor parameters do not allow phonation (both around -0.25 at the initialisation of the agent, line 1 of **Algorithm 1**) ; and there is no variations on the formant values, which makes the control task trivial as soon as the agent has a bit of experience. There is more to learn with *unarticulated* vocalizations, where formant values are varying in at least one part of the vocalization, and still more with *articulated* ones where they are varying in both parts (for the first and second command).

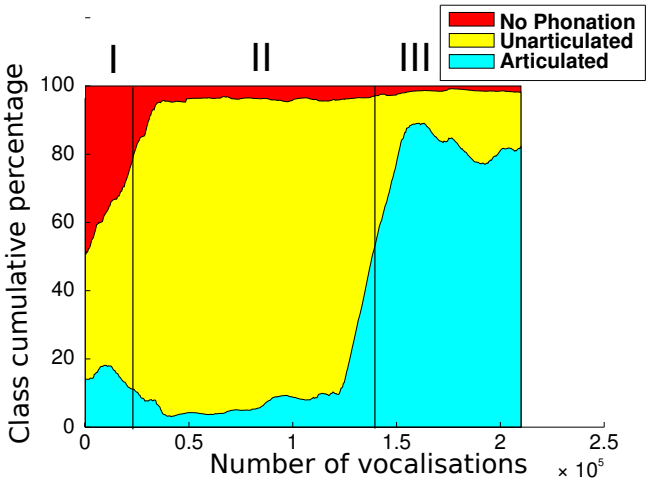

Figure 7. Self-organization of vocal developmental stages. At each time step t (x-axis), the percentage of each vocalization class between t and $t + 30.000$ is plotted (y-axis), in a cumulative manner (sum to 100%). Vocalization classes are defined in section 2.1.3. Roman numerals shows three distinct developmental stages. I: mainly no phonation or unarticulated vocalizations. II: mainly unarticulated. III: mainly articulated. The boundaries between these stages are not preprogrammed and are here manually set by the authors, looking at sharp transitions between relatively homogeneous phases.

Types of sounds produced	Stage I	Stage II	Stage III	Stage IV
No phonation-Unarticulated	**7**	0	2	0
Unarticulated	0	**7**	0	3
Articulated	0	2	**4**	0
Other	2	0	1	0

Table 1. Count of vocalization stages in the 9 simulations of the supplementary data. The "types of sounds produced" (first column of the table) correspond to the most prominent class in a given stage, where stages are manually set, looking at sharp transitions between relatively homogeneous phases. These developmental stages are therefore subjective to a certain extent, in the sense that another observer could have set different ones (but hopefully also would observe major structural changes). "No phonation-Unarticulated" means a mix between *No phonation* and *Unarticulated* classes (as defined in section 2.1.3 in that stage. A number x in a cell means *this type of vocalizations (row) appears x times at the n^{th} stage of development (column) in the set of 9 simulations*. Two to four developmental stages were identified in each simulation, explaining why the "Stage I" and "Stage II" columns sum up to 9 (the total number of simulations), but not the "Stage III" and "Stage IV" columns.

Figure 8 shows what happens in the particular simulation of **Figure 7** in more details.

This developmental sequence is divided into 3 stages, I, II and III, stages being separated by vertical dark lines on **Figure 8**, identical on each subplot (stage boundaries are the same than in **Figure 7**).

In stage I, until approximately 30.000 vocalizations, the agent produces mainly *no phonation* and *unarticulated* vocalizations. We observe that the agent set goals for $I(1)$ either around 0, either around 1, whereas the goals for $I(2)$ stay around 0 (last row in "Goals"). By trying to achieve these goals, the agent progressively refines its sensorimotor model and progresses by raising the values of the pressure and voicing motor parameter in the first command (two last rows of the section "Motor commands", 1st column). Other articulators remain around the neutral position (value 0). The agent is learning to phonate. The percentages of vocalization belonging to each vocalization class is provided **Table 2**.

Then, in stage II, from 30.000 to approximately 150.000 vocalizations, the agent is mainly interested in producing vocalizations which begin with a *Vowels* ($I(1) > 0.9$, see the definition of phone types in section 2.1.3) and finish with a *None* ($I(2) < 0.1$). An example of such a VN vocalization can be observed in the supplementary data, **Figure 14** in section 4.1. During this stage, it learns to produce relatively

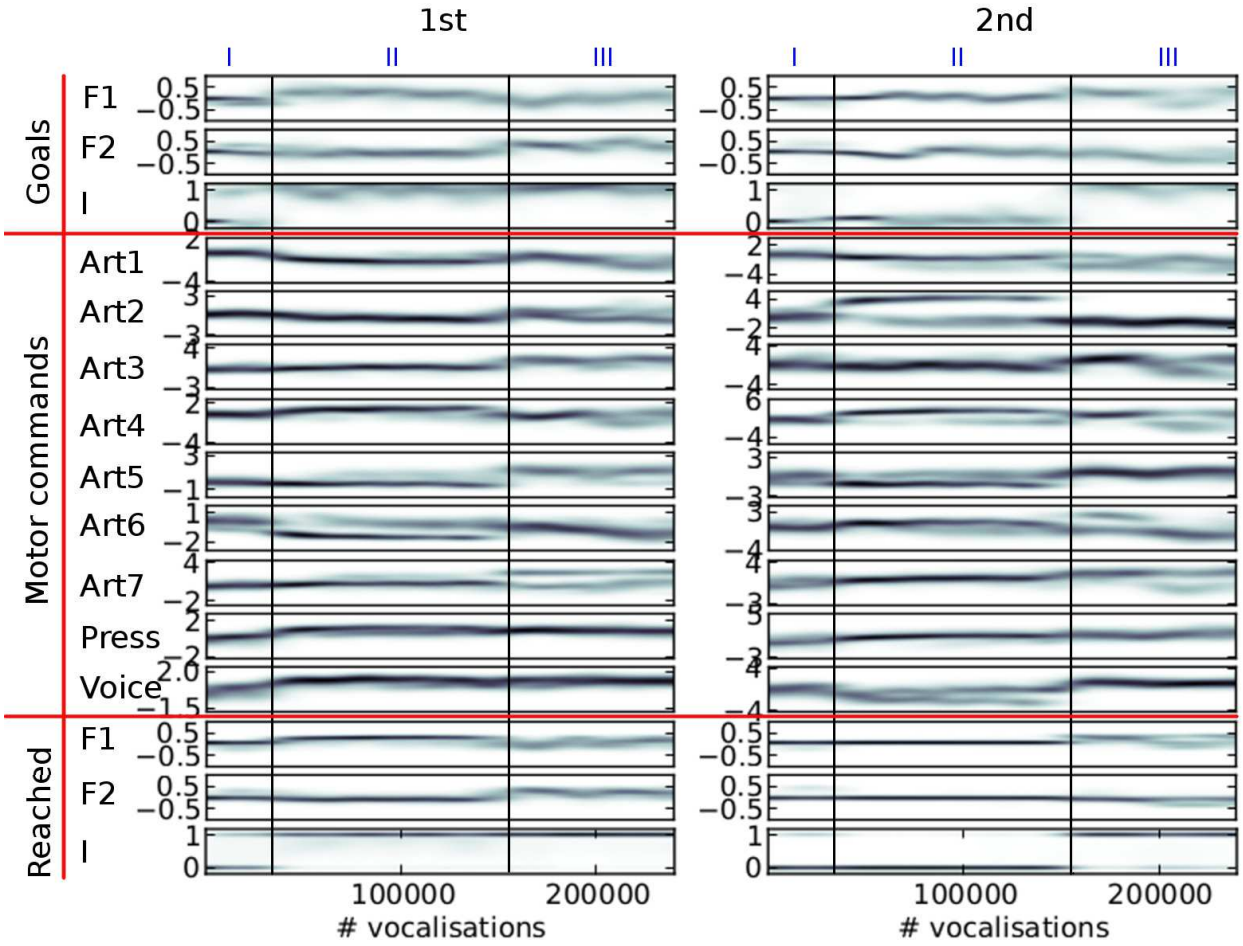

Figure 8. Evolution of the distribution of auditory goals, motor commands and sounds actually produced over the life time of a vocal agent (the same agent as in **Figure 7**). The variables are in three groups (horizontal red lines): the goals chosen by the agent in line 3 of **Algorithm 1** (top group), the motor commands it inferred to reach the goals using its inverse model in line 4 (middle group), and the actual perceptions resulting from the motor commands through the synthesizer in line 5 (bottom group). There are two columns (1st and 2nd), because of the sequential nature of vocalizations (two motor commands per vocalization). Each subplot shows the density of the values taken by each parameter (y-axis) over the life time of the agent (x-axis, in number of vocalizations since the start). It is computed using an histogram on the data (with 100 bins per axis), on which we apply a 3-bins wide Gaussian filter. The darker the color, the denser the data: e.g. the auditory parameter I actually reached by the second command ($I(2)$, last row in 'Reached", 2nd column), especially takes values around 0 (y-axis) until approximately 150.000^{th} vocalization (x-axis), then it takes rather values around 1. The three developmental stages of **Figure 7** are reported at the top.

high $F1(1)$ values, in particular by decreasing the $Art_1(1)$ parameter (approximately controlling the jaw height, see **Figure 3**). Regarding the second command, although the agent self-generates various goals for $F1(2)$ and $F2(2)$, and produces various motor commands to try to reach them, the sound produced mostly corresponds to a *None* ($I(2) = 0$, and therefore $F1(2) = F2(2) = 0$). This is due both to the negative value of the voicing parameter (last row in "Motor commands", second column), and to the fact that the vocal tract often ends in a closed configuration due to the poor quality of the sensorimotor model in this region (because phonation occurs very rarely for the second command, leaving the agent without an adequate learning set). During this stage, the agent explores a limited part of the sensorimotor space both in time (sound only for the first command) and space (around the neutral position), until it finally manages to phonate more globally at

NN	CN	NC	VN	NV	VV	CV	VC	CC
45.3 %	13.4 %	0.6 %	18.9 %	4.5 %	9.9 %	6.6 %	0.7 %	0.2 %

Table 2. Percentage of vocalization classes produced in stage I of the studied developmental sequence.

the end of this stage. This could be correlated to the acquisition of articulated vocalizations. The percentages of vocalization belonging to each vocalization class is provided in **Table 3**.

NN	CN	NC	VN	NV	VV	CV	VC	CC
4.0 %	26.9 %	0.1 %	62.2 %	0.1 %	3.4 %	0.5 %	2.5 %	0.2 %

Table 3. Percentage of vocalization classes produced in stage II of the studied developmental sequence.

Finally, in stage III (until 150.000 to the end), phonation almost always occurs during both the perception time windows (see I densities, both for goals and reached values). An example of such a VV vocalization can be observed in the supplementary data, **Figure 14** in section 4.1. This is much harder to achieve for two reasons: firstly because there is a need to control a sequence of 2 articulators movement in order to reach two formant values in sequence (i.e. $F1(1)$, $F1(2)$, $F2(1)$, $F2(2)$) instead of one in the previous stage (the second command leading to no sound), and secondly because the position of the articulators reached for the second command also depends on the position of the articulators reached for the first one (a kind of coarticulation due to the dynamical properties of the motor system). We observe that the range of values explored in the sensorimotor space is larger than for the previous stage (both in motor and auditory spaces). The percentages of vocalizations belonging to each vocalization class is provided in **Table 4**.

NN	CN	NC	VN	NV	VV	CV	VC	CC
1.6 %	3.7 %	0.1 %	12.1 %	0.8 %	67.5 %	6.5 %	6.8 %	0.8 %

Table 4. Percentage of vocalization classes produced in stage III of the studied developmental sequence.

3.2 INFLUENCE OF THE AUDITORY ENVIRONMENT

In a second set of experiments, we integrated a social environment providing a set of adult vocalizations. As explained in section 2.4, the learner has an additional choice: it can explore autonomously, or emulate the adult vocalizations. An "ambient language" is here modeled as a set of two speech sounds. To make it coherent with human language and the learning process observed in development, we chose speech-like sounds, typically vowel or consonant-vowel sounds. In terms of our sensorimotor descriptions, the adult sounds correspond to $I1$ with low values and $I2$ with high values. **Figure 9** shows such vocalizations corresponding to those used by Teacher 1 in **Figure 10** .

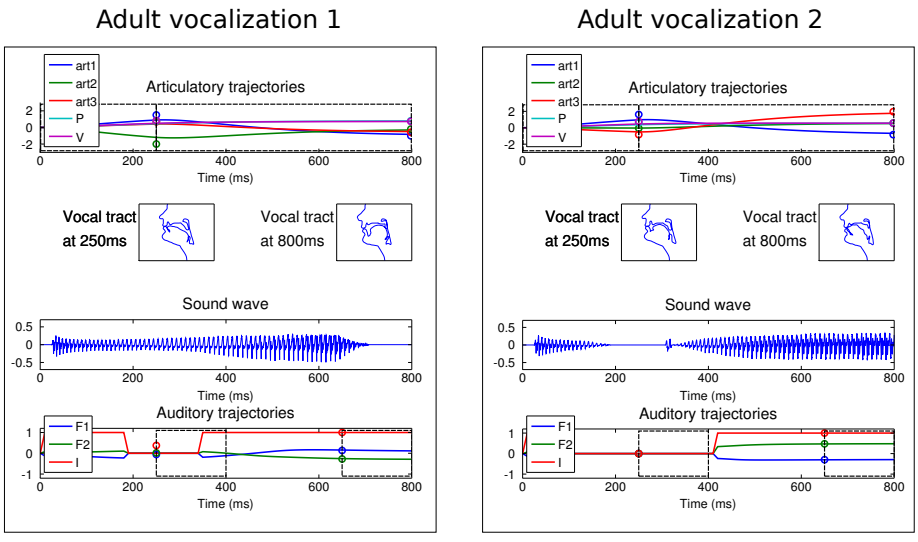

Figure 9. The two vocalizations of the adult Teacher 1 used in **Figure 10** , with the same convention as in **Figure 4** .

Figure 10 shows a significant evolution in the agent's vocalizations. In the early stage of its development, it can only make a few sounds. Most sounds correspond to small values of $I1(2), F1(1), F1(2), F2(1)$ and $F2(2)$, as in the first developmental stage of the previous

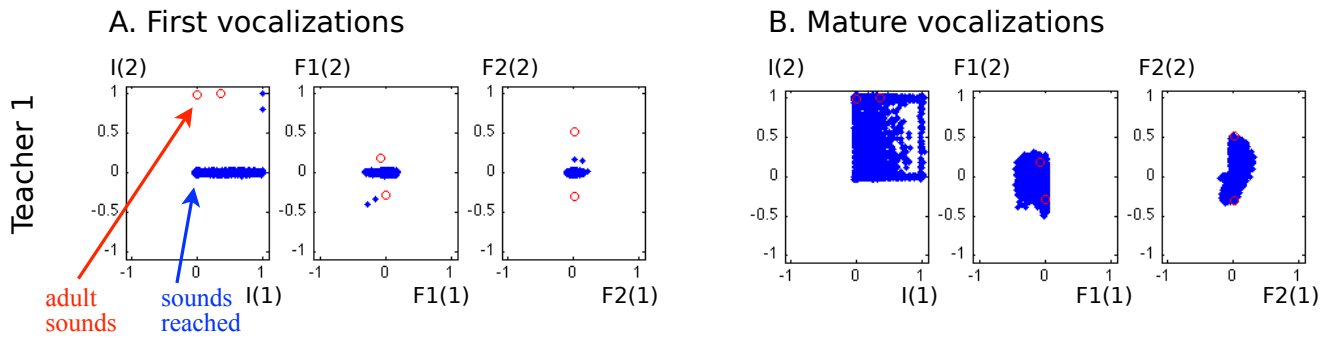

Figure 10. Vocalizations of the learning agent in the early and mature stages of vocal development. A) All auditory outcomes s produced by the agent in its early stage of vocalization are represented by blue dots in the 6-dimensional space of the auditory outcomes. The adult sounds are represented in red circles. The actually produced auditory outcomes only cover a small area of physically possible auditory outcomes, and correspond mostly to $I(2) = 0$, which represent vowel-consonant or consonant-consonant types of syllables. B) The auditory outcomes produced by the infant in its mature stage of vocalization cover a much larger area of auditory outcomes and extend in particular over areas in which vocalizations of the social peer are located.

experiment (see **Table 2** and **Figure 8**). Therefore the agent is not able to reproduce the ambient sounds of its environment. In contrast, in later periods of its development, its vocalizations cover a wider range of sounds, with notably $I(1)$ and $I(2)$ both positive, which means it now produces more articulated sounds. The development of vocalizations for a self-exploring agent in the last section showed that it progressively was able to produce articulated vocalizations, which we observed at times at the end of its development. This effect has been reinforced by the environment: with articulated vocalizations to emulate, it produces this class more regularly.

Another important result is that mature vocalizations can now reproduce the ambient sounds of the environment: the regions of the sounds produced by the learner (blue dots) overlap the teacher's demonstrations (red circles). It seems that, during the first vocalizations, the agent cannot emulate the ambient sounds because they are too far away from its possible productions, and thus it can hardly make any progress and approach these demonstrations. **Figure 11** confirms this interpretation. In the beginning, the agent makes no progress with emulation, and it is only around $t = 450$ that it makes progress with the emulation strategy. At that point, as we can see in **Figure 12**, it uses equally both strategies. This enables the agent to make considerable progress from $t = 450$ to $t = 800$. Indeed, once its mastery improves and the set of sounds it can produce increases, it then increasingly emulates ambient sounds. Once it manages to emulate the ambient sounds well, and thus its competence progress decreases, it uses less the emulation strategy and more the self-exploration strategy.

To analyse better this emulation phenomenon and assess the influence of the ambient language, we run the same experiment with different acoustic environments. We used two other sets of speech sound demonstrations from simulated peers, and analysed the auditory productions of the agent in **Figure 13**. The first property that can be noted is that in the early phase of the vocal exploration (**Figure 13. A** and **C**), the auditory productions of the two agents are alike, and do not depend on the speech environment. On the contrary, the mature vocalizations vary with respect to the speech environment. With Teacher 1, the productions have their values $F2(1)$ and $F2(2)$ along the axis formed by the demonstration (**Figure 10. A**, last column). Comparatively, Teacher 2's speech sounds have different $F1(1), F1(2), F2(1)$ and $F2(2)$. As represented in **Figure 13. B**, the two speech sounds now differ mainly by their $F1(1)$ (instead of $F1(2)$) and in their subspace $(F2(1), F2(2))$ the speech sounds have approximately rotated from those of Teacher 1. The produced auditory outcomes of the learner look like they have changed in the same way. Whereas the reached space (blue area) seemed to be along axis F1(2) and F2(2) and little on F1(1) or F2(1) for Teacher 1, it has extended its exploration along F1(2) and F2(2) for Teacher 2. With Teacher 3, the demonstrations are more localised in the auditory space, with $F1(1) < 0$ and $F2(2) > 0$. The effect we observe in **Figure 13. D** is that the exploration is more localised too: the explored space is more oriented toward areas where $F1(1) < 0$ and $F2(2) > 0$. Thus, these three examples strongly suggest a progressive influence of the auditory environment, in the sense that the first vocalizations in **Figure 10** and **13** are very similar, whereas we observe a clear influence of the speech environment on the produced vocalizations in later stages.

Altogether, the results of these experiments provide a computational support to the hypothesis that the progressive influence of the ambient language observed in infant vocalizations can be driven by an intrinsic motivation to maximize competence progress. At early developmental stages, attempts to imitate adult vocalizations are certainly largely unsuccessful because basic speech principles, such as phonation, are not yet mastered. In this case, focusing on simpler goals probably yields better progress niches than an imitative behavior. While they are progressively mastered, the interest in these goals decreases whereas the ability to imitate adult vocalizations increases. Imitation thus becomes a new progress niche to explore.

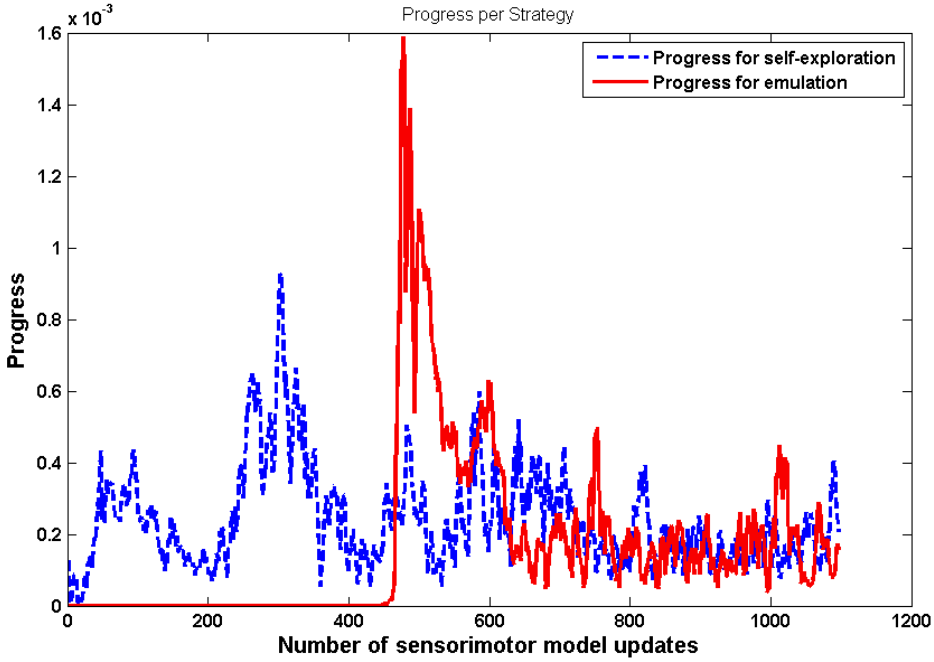

Figure 11. Progress made by each strategy with respect to the number of updates of the sensorimotor model G_{SM}. These values have been smoothened over a window of 100 updates. For $t < 450$, the agent makes no progress using emulation strategy. After $t = 450$, both strategies enable the agent to make progress.

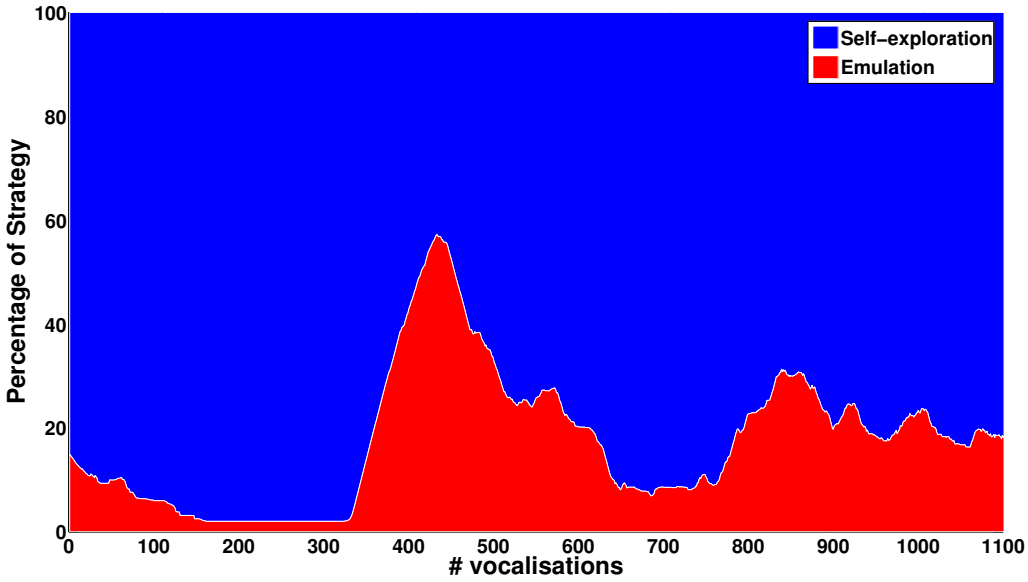

Figure 12. Percentage of times each strategy is chosen with respect to the number of updates of the sensorimotor model G_{SM}. These values have been smoothened over a window of 100 updates. For $t < 450$, the agent mainly uses self-exploration strategy. When its knowledge enables it to make progress in emulation, it chooses emulation strategy until it can emulate the ambient sounds well (and its competence progress decreases).

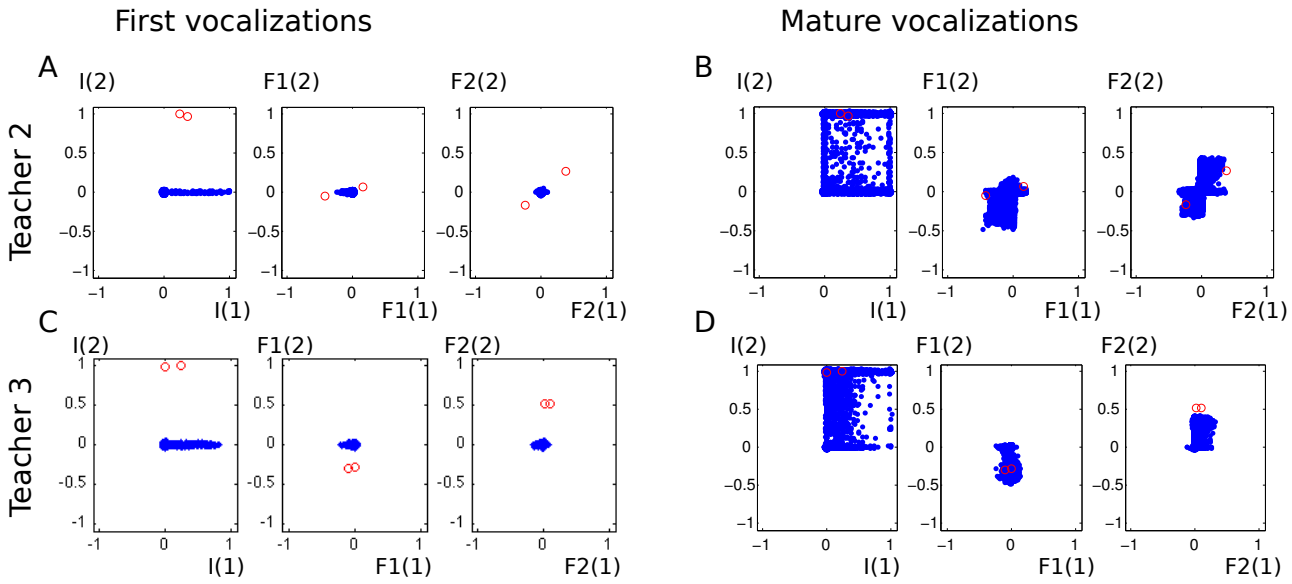

Figure 13. Vocalizations of the learning agent in the early and mature stage of vocalization in two different speech environments (Teacher 2 and Teacher 3). A and C) All auditory outcomes produced by the vocal learner in its early stage of vocal development are represented by blue dots in the 6-dimensional space of the auditory outcomes. The sounds of the environment are represented in red circles. The auditory outcomes only cover a small area, and do not depend on the speech environment. B and D) The auditory outcomes produced by the infant in its mature stage of vocal development cover a larger area of auditory outcome, which depend on the speech environment.

4 DISCUSSION

Our main contribution with respect to previous computational models of speech acquisition is that we do not presuppose the existence of successive developmental stages, but rather they can emerge from an intrinsic drive to maximize the competence progress. We showed that vocal developmental stages can self-organize autonomously, from simple sensorimotor activities to more complex ones. The agent starts producing *no phonation* and *unarticulated* vocalizations, which are easy to produce because limited in the range of their auditory effects. This can be related to the first stage in infant vocal development (**Figure 1**), where the agent produces non speech-sounds (e.g. growls, squeals...) before learning phonation and then produces not well-articulated quasi-vowels. Later on, once the agent does not progress much in producing *unarticulated* vocalizations, it focuses on more complex vocalizations of the *articulated* class. The reason is that, due to the properties of the sensorimotor system and internal model, the mastering of complex tasks require first the mastering of simpler tasks in order to yield significant competence progress, so that these complex tasks are selected as interesting goals.

We also showed that intrinsically motivated exploration can lead to a progressive interest towards the sounds of the ambient language. Whereas the first vocalizations are mainly the result of self-exploration, they progressively lead to mastering necessary speech principles (e.g. phonation). This progressive mastering allows in turn to make significant progress in adult-speech imitation, which explains why the vocal learner starts to choose more often as targets the sound of its environment. Competence-progress based curiosity-driven exploration could thus explain a progressive influence of the ambient language on infant vocalizations.

We therefore showed that intrinsically motivated active exploration can self-organize a coherent developmental sequence, without any external clock or preset specification of this sequence. This possible role of intrinsic motivation, providing a mechanism to discover autonomously necessary developmental stages to structure the learning process, is here validated computationally. We believe that it could be of major interest for understanding the structuration of early vocal development in infants. Speech acquisition is such a complex task that intrinsic motivation could be a crucial component to make it possible in the infant's first year of life.

Our model, however, has a number of limitations. Firstly, our modeling choices of the articulatory and auditory representations, as well as the implementation of the transformation from the former to the latter, is somewhat less realistic than in some previous models: articulatory trajectories are specified using two commands per articulator with fixed durations and the auditory representation uses only three acoustic parameters (the intensity and the two first formants) averaged in fixed and relatively arbitrary perception time windows. Moreover, the fact that formant values are set to 0 whenever the intensity of the signal is null can appear quite unrealistic. Although previous models often provide more meticulous implementations of the sensorimotor system, including e.g. pitch or tactile information, what is important to us is a sensorimotor system where all vocalizations are not equally easy to learn in terms of control. Such a requirement is certainly necessary for a clear developmental sequence to emerge. Secondly, we did not treat a major issue in speech acquisition research, the so-called correspondence problem: how the child is able to relate its own vocalizations to adult vocalizations, whereas the vocal tract of the child is very different in size and geometry than the one of an adult, and therefore the spectral characteristics of the produced sounds are different. Solutions to overcome this problem have been proposed, generally based on adult feedback or reformulations associated with infant productions [60, 61, 13]. This is outside the scope of this paper where our focus is on the self-organization of the developmental sequence in successive stages of increasing complexity. Extending our model to the interaction with real humans would definitely require to consider this issue.

Further works will consider higher-dimensional sensorimotor spaces for more realism. For example, the free software Praat [62] is a powerful tool allowing to synthesize a speech signal from a trajectory in a 29-dimensional space of respiratory and oro-facial muscles. Numerous acoustic features can in turn be extracted from the synthesized sound, among which the Mel-frequency cepstral coefficients (MFCC, [63]). It would also be interesting to study the effect of a vocal tract growing during the learning process, to study if our intrinsically motivated agent could re-explore only parts of the sensorimotor space which were the most affected by the vocal tract shape change. Generally, we believe that a developmental robotics approach applied to a realistic articulatory model can appropriately manage the learning process of a complex and changing mapping in high-dimensional spaces, and that observed developmental sequences can lead to interesting comparisons with infant data and predictions. Regarding the present study, such a prediction could be that a human infant should be influenced by adult sounds earlier if they were easier to produce than well-formed syllables. For example, one could imagine an experiment in which a very young infant is put in an environment where he hears external sounds that are simpler than vowels/consonants/syllables (e.g. growls) and test whether his vocalizations become influenced by external environment earlier and/or if we can measure a greater interest than in a normal speech environment.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGEMENT

The authors would like to thank Louis-Jean Boë for the design of Figure 2 (vocal tract by Sophie Jacopin).

Funding: This work was partially financed by ERC Starting Grant EXPLORERS 240 007.

SUPPLEMENTARY DATA

4.1 VOCALIZATION TYPES

Figure 14 shows the 9 types of vocalizations defined in section 2.1.3 (NN, CN, NC, VN, NV, VV, VC, CV and CC).

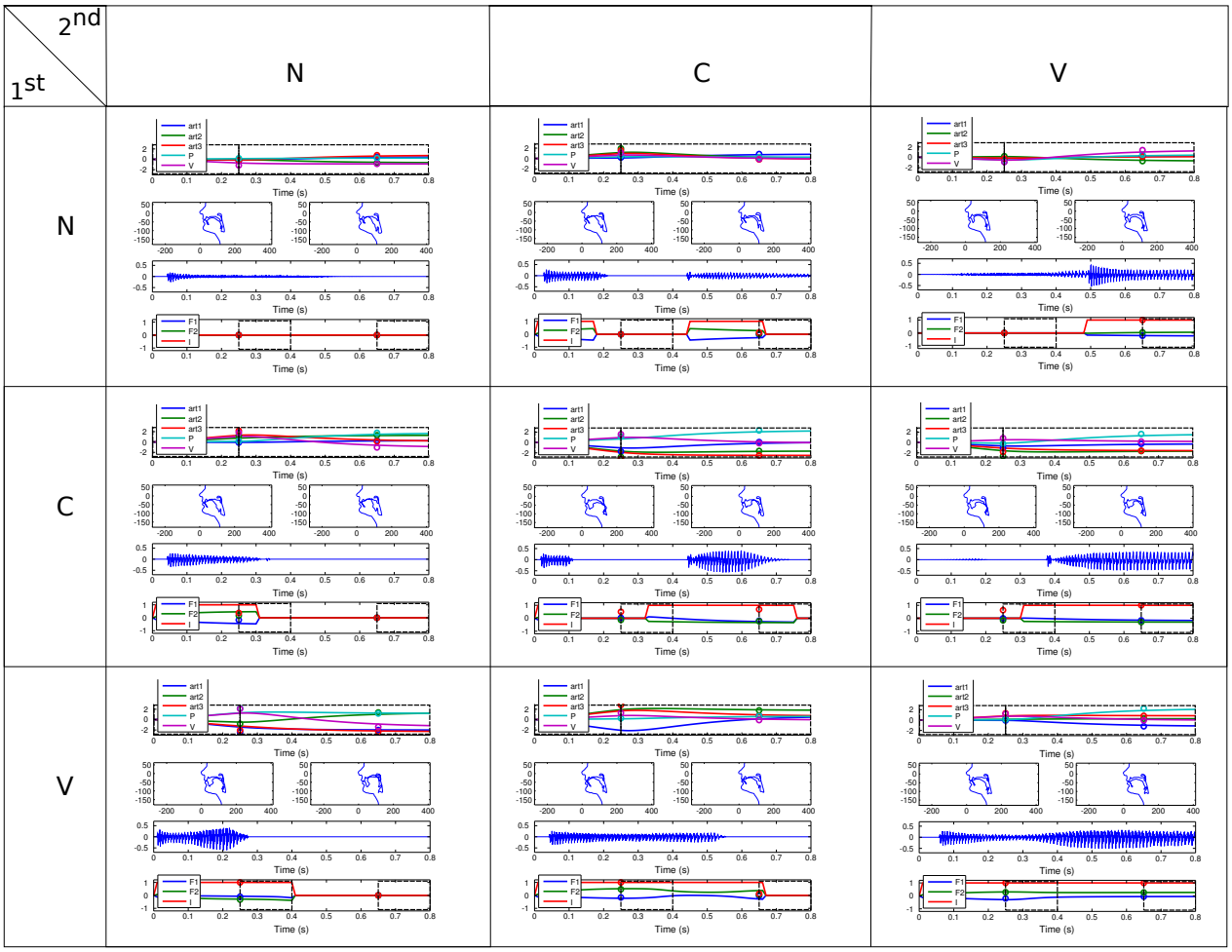

Figure 14. Examples of each vocalization types. Rows (1st) correspond to the type of the first phone and columns (2nd) to the type of the second phone of the vocalization. There are three possible phone types, as defined in section 2.1.3: the *Vowels* (V) which have a high intensity ($I > 0.9$), the *Consonants* (C) which have a low intensity ($0.1 < I < 0.9$) and the *None* which have almost no intensity ($I < 0.1$). For example, the plot in the second row (C) third column (V) corresponds to a CV vocalization, with the same convention as in **Figure 4**.

4.2 DEVELOPMENTAL SEQUENCES OF 9 INDEPENDENT SIMULATIONS

The figures of this section display the emerging developmental sequence of 9 independent simulations in pure self-exploration mode (section 3.1). At each time step t (x-axis), the percentage of each vocalization class during between t and $t + 30.000$ is plotted (y-axis), in a cumulative manner. Vocalization classes are defined in section 2.1.3. For each one, we show boundaries between developmental stages. These boundaries are set manually, by looking at sharp transitions between relatively homogeneous phases. They are therefore subjective to a certain extent, in the sense that another observer could have set different ones (but hopefully also would observe major structural changes).

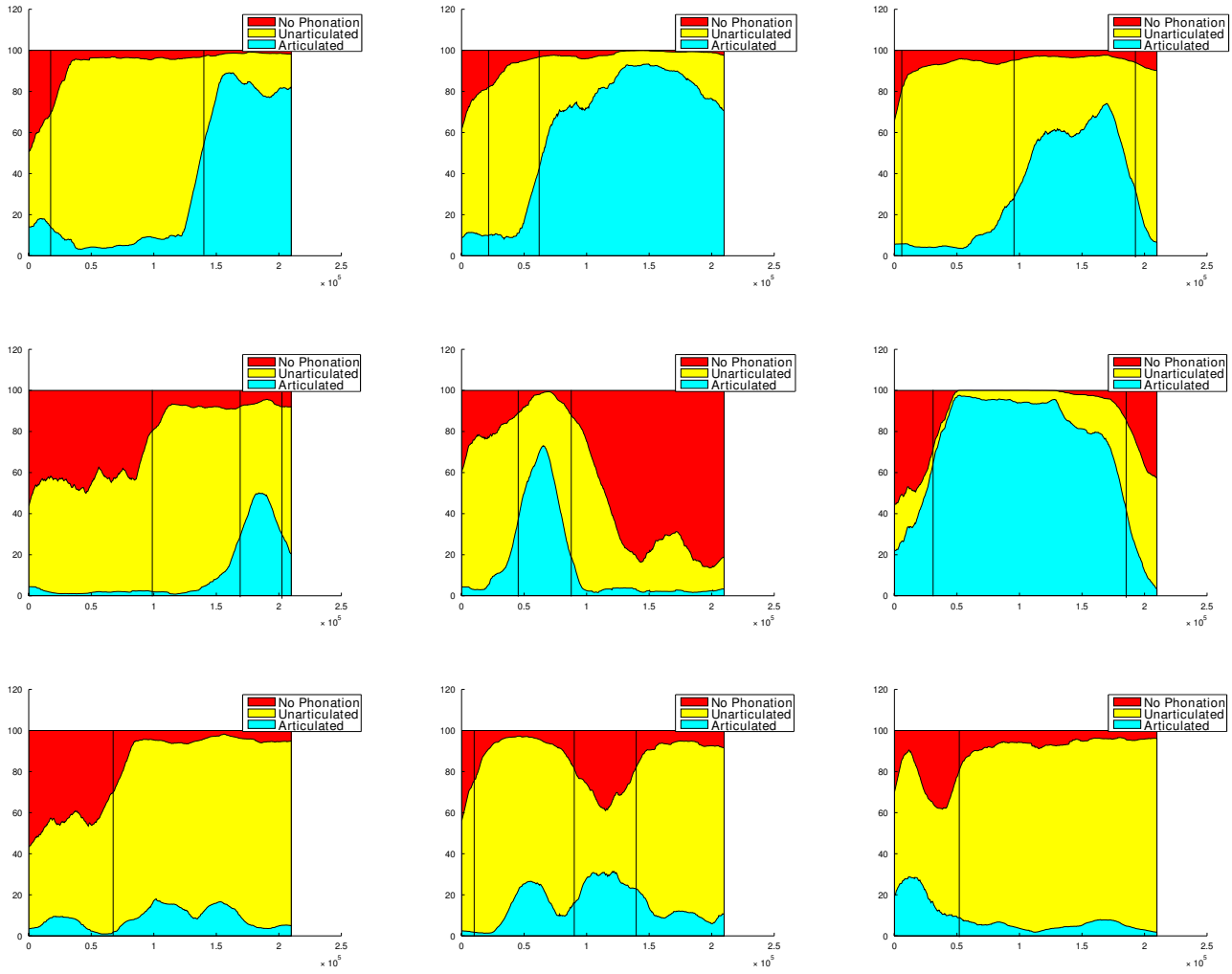

Figure 15. Developmental sequences emerging from the 9 simulations for the experiment described in section 3.1. Each subplot follows the same convention as in **Figure 7**. The simulations have been ordered, also in a subjective manner, from those which display a clear developmental sequence of the type *No phonation* → *Unarticulated* → *Articulated* to those less organized (from left to right, then top to bottom).

REFERENCES

[1]D. Kimbrough Oller. *The Emergence of the Speech Capacity*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[2]B. Sigismund. *Child language: a book of readings*, chapter Kind und Welt, pages 17–18. Englewood Cliffs, NJ: Prentice-Hall. (Original work published in 1856), 1971.

[3]H. Taine. *Child language: a book of readings*, chapter Acquisition of language by children, pages 20–26. Englewood Cliffs, NJ: Prentice-Hall. (Original work published in 1856), 1971.

[4]D. E. Berlyne. A theory of human curiosity. *British Journal of Psychology*, 45:180–191, 1954.

[5]E.L. Deci and Richard M. Ryan. *Intrinsic Motivation and self-determination in human behavior*. Plenum Press, New York, 1985.

[6]Richard M Ryan and Edward L Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000.

[7]M. Csikszentmihalyi. *Creativity: Flow and the Psychology of Discovery and Invention*. HarperCollins, 1997.

[8]Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593, 2013.

[9]Karl Friston, Rick A Adams, Laurent Perrinet, and Michael Breakspear. Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology*, 3, 2012.

[10]F H Guenther, M Hampson, and D Johnson. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, 105(4):611–633, October 1998. PMID: 9830375.

[11]Frank H. Guenther. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5):350–365, September 2006.

[12]Bernd J Kröger, Jim Kannampuzha, and Christiane Neuschaefer-Rube. Towards a neurocomputational model of speech production and perception. *Speech Communication*, 51(9):793–809, 2009.

[13]I.S. Howard and P. Messum. Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, 15(1):85–117, 2011.

[14]A. S Warlaumont. Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, 38:64–95, 2013.

[15]A.S. Warlaumont. A spiking neural network model of canonical babbling development. In *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*, pages 1–6, 2012.

[16]A.S. Warlaumont. Salience-based reinforcement of a spiking neural network leads to increased syllable production. In *Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE International Conference on*, pages 1–6, 2013.

[17]Pierre-Yves Oudeyer, Frederic Kaplan, and Verena Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286, 2007.

[18]Adrien Baranes and Pierre-Yves Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.

[19]Pierre-Yves Oudeyer, Adrien Baranes, Frederic Kaplan, and Olivier Ly. *Intrinsically Motivated Learning in Natural and Artificial Systems*, chapter Developmental constraints on intrinsically motivated skill learning: towards addressing high-dimensions and unboundedness in the real world. Springer, 2013.

[20]J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In J. A. Meyer and S. W. Wilson, editors, *Proc. SAB'91*, pages 222–227, 1991.

[21]A Barto, S Singh, and N. Chenatez. Intrinsically motivated learning of hierarchical collections of skills. In *ICDL International Conference on Developmental Learning*, pages 112–119, San Diego, CA, 2004.

[22]Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 2007.

[23]G. Baldassarre. What are intrinsic motivations? a biological perspective. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, pages 1–8. IEEE, 2011.

[24]Gianluca Baldassarre and Marco Mirolli. *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer, 2013.

[25]Rupesh Kumar Srivastava, Bas R Steunebrink, and Jürgen Schmidhuber. First experiments with powerplay. *Neural Networks*, 41:130–136, 2013.

[26]J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[27]M Schembri, M Mirolli, and G Baldassarre. Evolving childhood's length and learning parameters in an intrinsically motivated reinforcement learning robot. In *Proceedings of the seventh international conference on epigenetic robotics*, volume 134, pages 141–148. Lund: Lund University, 2007.

[28]Kathryn Merrick and Mary Lou Maher. Motivated learning from interesting events: adaptive, multitask learning agents for complex environments. *Adaptive Behavior*, 17(1):7–27, 2009.

[29]S. Hart. An intrinsic reward for affordance exploration. In *ICDL International Conference on Developmental Learning*, 06 2009.

[30]Andrew Stout and Andrew G Barto. Competence progress intrinsic motivation. In *Development and Learning (ICDL), 2010 IEEE 9th International Conference on*, pages 257–262. IEEE, 2010.

[31]A. Baranes and P-Y. Oudeyer. Intrinsically motivated goal exploration for active motor learning in robots: a case study. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010), Taipei, Taiwan*, 2010.

[32]F. Kaplan and P-Y. Oudeyer. The progress-drive hypothesis: an interpretation of early imitation. In K. Dautenhahn and C. Nehaniv, editors, *Models and mechanisms of imitation and social learning: Behavioural, social and communication dimensions*. Cambridge University Press, 2007.

[33]Pierre-Yves Oudeyer and Frederic Kaplan. Discovering communication. *Connection Science*, 18(2):189–206, 06 2006.

[34]Clément Moulin-Frier and Pierre-Yves Oudeyer. Curiosity-driven phonetic learning. In *International Conference on Development and Learning, Epirob, San Diego, USA*, 2012.

[35]Frederic Kaplan and Pierre-Yves Oudeyer. In search of the neural circuits of intrinsic motivation. *Frontiers in neuroscience*, 1(1):225, 2007.

[36]Sao Mai Nguyen and Pierre-Yves Oudeyer. Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner. *Paladyn Journal of Behavioural Robotics*, 3(3):136–146, 2012.

[37]Manuel Lopes and Pierre-Yves Oudeyer. The strategic student approach for life-long exploration and learning. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–8. IEEE, 2012.

[38]Clément Moulin-Frier and Pierre-Yves Oudeyer. The role of intrinsic motivations in learning sensorimotor vocal mappings: a developmental robotics study. In *Proceedings of Interspeech*, page In press, Lyon, France, 2013.

[39]Clément Moulin-Frier and Pierre-Yves Oudeyer. Exploration strategies in developmental robotics: a unified probabilistic framework. In *International Conference on Development and Learning, Epirob, Osaka, Japan*, 2013.

[40]E. Thelen and L.B. Smith. *A Dynamic Systems Approach to the Development of Cognition and Action*. A Bradford book. MIT Press, 1996.

[41]Gilbert Gottlieb. Experiential canalization of behavioral development: Theory. *Developmental Psychology*, 27(1):4, 1991.

[42]Marilyn M Vihman, Charles A Ferguson, and Mary Elbert. Phonological development from babbling to speech: Common tendencies and individual differences. *Applied Psycholinguistics*, 7(1):3–40, 1986.

[43]P. K. Kuhl. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5(11):831–843, 2004.

[44]Frank H Guenther, Satrajit S Ghosh, and Jason A Tourville. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, 96(3):280–301, 2006.

[45]S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. *Speech production and speech modelling*, 55:131–149, 1989.

[46]Kevin Lee Markey. *The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development*. PhD thesis, University of Colorado at Boulder, 1994.

[47]Paul Boersma. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Holland Academic Graphics, 1998.

[48]A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[49]Sylvain Calinon. *Robot Programming by Demonstration: A Probabilistic Approach*. CRC, 2009.

[50]Manuel Lopes, Francisco Melo, Luis Montesano, and Jos Santos-Victor. Abstraction levels for robotic imitation: Overview and computational approaches. In Olivier Sigaud and Jan Peters, editors, *From Motor Learning to Interaction Learning in Robots*, volume 264 of *Studies in Computational Intelligence*, pages 313–355. Springer Berlin Heidelberg, 2010.

[51]J. Call and M. Carpenter. *Imitation in animals and artifacts*, chapter Three sources of information in social learning, pages 211–228. Cambridge, MA: MIT Press., 2002.

[52]Andrew Whiten. Primate culture and social learning. *Cognitive Science*, 24(3):477–508, 2000.

[53]Chrystopher L Nehaniv and Kerstin Dautenhahn. *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*. Cambridge Univ. Press, Cambridge, March 2007.

[54]Paul Van Geert. A dynamic systems model of cognitive and language growth. *Psychological review*, 98(1):3, 1991.

[55]Linda B. Smith and Esther Thelen. Development as a dynamic system. *Trends in Cognitive Sciences*, 7(8):343 – 348, 2003.

[56]Ian Maddieson and Kristin Precoda. Updating UPSID. *The Journal of the Acoustical Society of America*, 86(S1):S19, November 1989.

[57]J.-L. Schwartz, L.-J. Boë, N. Vallée, and C. Abry. Major trends in vowel system inventories. *Journal of Phonetics*, 25(3):233–253, 1997.

[58]Clément Moulin-Frier, Jean-Luc Schwartz, Julien Diard, and Pierre Bessière. *Primate communication and human language: Vocalisations, gestures, imitation and deixis in humans and non-humans*, chapter Emergence of articulatory-acoustic systems from deictic interaction games in a "Vocalize to Localize" framework, pages 193–220. Advances in Interaction Studies' series by John Benjamins Pub. Co., 2011.

[59]Pierre-Yves Oudeyer. The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3):435–449, April 2005.

[60]Hisashi Ishihara, Yuichiro Yoshikawa, Katsushi Miura, and Minoru Asada. How caregiver's anticipation shapes infant's vowel through mutual imitation. *Autonomous Mental Development, IEEE Transactions on*, 1(4):217–225, 2009.

[61]Katsushi Miura, Yuichiro Yoshikawa, and Minoru Asada. Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver. *Advanced Robotics*, 26(1-2):23–44, 2012.

[62]David (2012). Boersma, Paul & Weenink. Praat: doing phonetics by computer [computer program]. `http://www.praat.org/`, 2012.

[63]S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, August 1980.